IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant:	Keith V. Wood et al.	Examiner: Rebecca E. Prouty
Serial No.:	09/645706	Group Art Unit: 1652
Filed:	August 24, 2000	Docket No.: 341.005US1
Title:	SYNTHETIC NUCLEIC ACID MOLECULE COMPOSITIONS AND METHODS OF PREPARATION	

DECLARATION UNDER 37 C.F.R. § 1.132

Commissioner for Patents
Washington, D.C. 20231

Sir:

I, Monika Wood, M.S., declare and say as follows:

1. I am one of the named co-inventors of the claims in the above-identified application. I make this Declaration in support of the patentability of the claims of the above-identified application.

2. Sherf et al. (U.S. Patent No. 5,670,356) disclose that a firefly luciferase (*luc*) gene was modified using mammalian codon replacement to remove 3 internal palindromic sequences, 5 restriction endonuclease sites, 4 glycosylation sites, and 6 transcription factor binding sites, yielding *luc*+. On June 14th 2006, using publicly available software and a database of transcription factor binding sites (see attached details on the specific software, search parameters, and database release used), comparable to those employed in the above-referenced application, potential mammalian transcription factor binding sites were identified in the *luc*+ gene. I found that the *luc*+ gene contains over 150 potential mammalian transcription factor binding sites.

3. Thus, mammalian transcription factor binding sites in a particular nucleic acid sequence can be identified and enumerated.

4. I further declare that all statements made herein of my own knowledge are true, and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title

18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Dated: June-19-2006 By: M Wood
Monika Wood

TESS - Filtered String Search Page

Check our FAQ page then please send questions and comments to TessMaster@cbil.upenn.edu.

Database Versions: TRANSFAC=4.0, IMD=v1.1, CBIL/GibbsMat=v1.1

New Feature: The tabular display page of the site search now displays the ordinal of the sorted hits.

The site is basically working. Please report any errors you encounter.

To keep the load on our server to a reasonable level, we have implemented a cap on the number of jobs that are waiting to execute. When submitting a search job, you may see a message asking you to submit your job later. In that case, wait a few minutes and try again.

What potential transcription factor binding sites are there in my sequence?

Input

Enter the minimal information needed to submit a job to TESS.

Title: luc+ *(text)*

You may submit multiple sequences in this window. Each sequence can have a maximum length of 2000[bp]. The sequences must a total length less than 100000[bp].

DNA Sequence(s):

```
atggaagacgccaaaaacataaagaaaggcccggcgccattctatccgctggaagatgga
accgctggagagcaactgcataaggctatgaagagatacgccctggttcctggaacaatt
gcttttacagatgcacatatcgaggtggacatcacttacgctgagtacttcgaaatgtcc
gttcggttggcagaagctatgaaacgatatgggctgaatacaaatcacagaatcgtcgta
tgcagtgaaaactctcttcaattctttatgccggtgttgggcgcgttatttatcggagtt
gcagttgcgcccgcgaacgacatttataatgaacgtgaattgctcaacagtatgggcatt
tcgcagcctaccgtggtgttcgtttccaaaaaggggttgcaaaaaattttgaacgtgcaa
aaaaagctcccaatcatccaaaaaattattatcatggattctaaaacggattaccaggga
tttcagtcgatgtacacgttcgtcacatctcatctacctcccggttttaatgaatacgat
tttgtgccagagtccttcgatagggacaagacaattgcactgatcatgaactcctctgga
```

(text)

Length of time to store results of the job: day

Your email address: *(string)*

End of Minimal Parameters

You can click 'Submit' to submit the job or scroll down to change the basic search parameters.

Submit

Databases

Check off the databases you want to include in the search and enter your own search strings and/or weight matrices.

String Databases

Search TRANSFAC Strings ☑

Search *My* Site Strings: ☐

(*text*)

Factor Filters

Use this section to control which factors are included in the search.

The buttons control the number of terms used in the filter. When you click on a button this form will reload with the adjusted number of terms.

Fewer | More | 0 | 1 | 2 | 3 | 4 | 5

Factor Attribute 1: Organism Classification

matches: mammalia (*text*)

Score Filters

Adjust these parameters to control the required strength of the match between the site and the string or weight matrix model.

String Scoring

Use these parameters if you have chosen to search for string matches.

Use only core positions for TRANSFAC strings: ☐

Maximum Allowable String Mismatch % (t_{mm}): 0

Minimum log-likelihood ratio score (t_{s-a}): 10 (*real number>=0)*

Minimum string length (t_w): 5

Output Control

Secondary Lg-Likelihood Deficit: 1.6 (*real number>=0.0 and <=6.0)*

☐ Count significance threshold: 1.0e-2 (*real number>=0.0 and <=1.0)*

Click Submit or scroll down to adjust the expert search parameters.

Submit

Submit - Reset

Help

Title:

This is a short title for your sequence which will appear in the results. Using as you see fit to identify the sequence.

The parameter's type is text. An item of text comprises an arbitrary sequence of characters, possibly including white-space and newlines.

DNA Sequence(s):

Enter your nucleic acid sequence(s) here using the IUB standard. TESS ignores the case of the letters and the presence of digits and white space. That means you can cut the sequence section from a GenBank entry and paste in here without any editing.

You may submit multiple sequences in this window. They will be processed individually. For example:

```
>Seq1
acgtagtagagctaga
>Seq2
acgtagcatgactgggatatatatatat
```

The parameter's type is text. An item of text comprises an arbitrary sequence of characters, possibly including white-space and newlines.

Length of time to store results of the job:

Select the length of time you want us to store the results of this job. We'll try to keep it this long.

The parameter's type is text. An item of text comprises an arbitrary sequence of characters, possibly including white-space and newlines.

Your email address:

Enter the email address to which you want material sent.

The parameter's type is string. A string is a single group of non-white space characters, e.g. 'this-is-a-string' but not 'this is not'.

Search TRANSFAC Strings:

Select this option to search for matches to TRASNSFAC sites in your query sequence.

Search *My* Site Strings:

Select this option and enter site strings if you want search for your own site strings in the query sequence(s).

Strings should be placed one per line with a trailing name separated from the sequence by one or more spaces or tabs. They will be assigned accession numbers from the series U00001, U00002, etc.

Here is a sample:

```
agtctgannnnagtca  factor x
aggtggaa          hairy eyeball
```

The parameter's type is text. An item of text comprises an arbitrary sequence of characters, possibly including white-space and newlines.

Factor Attribute 1:

Choose an attribute of the FACTOR database which you want to use to select factors to search for in your sequence.

See also: rxp

The parameter's type is text. An item of text comprises an arbitrary sequence of characters, possibly including white-space and newlines.

matches:

Enter the pattern that must be found in a factor's attribute to be included in the search. You can

follow these links to see what possible values each field takes on. This may help in forming queries.

- Organism Species
- Organism Classification
- Name
- Synonyms
- Interacting Factors
- Class
- Cell Positive Specificity
- Cell Negative Specificity
- Id

Use the "External Database References" option to search for factors by their EMBL, SwissProt, Flybase, Compel, or PIR accession numbers or ids. For example to find 'EMBL: J03236; MMJUNBA' you can enter either 'J03236' or 'MMJUNBA'.

See also: att

The parameter's type is text. An item of text comprises an arbitrary sequence of characters, possibly including white-space and newlines.

Use only core positions for TRANSFAC strings:

Site strings in TRANSFAC indicate which positions are important to binding but also include unimportant positions as well.

If you select this option, then the unimportant positions will be removed from the site string prior to searching.

The parameter's type is Boolean. Either True/False, Yes/No, 0/1, or On/Off.

Maximum Allowable String Mismatch % (t_{mm}):

TESS will consider a transcription element to match a part of your sequence as long as the percentages of mismatch is below the specified level.

The number you enter here is an integer percent value and so must be between 0 and 100.

The parameter's type is integer. A series of digits optionally preceeded by a minus sign. Commas are ok.

Minimum log-likelihood ratio score (t_{s-a}):

TESS will not report string matches with a log-likelihood ratio less than this value. Use this value as a hedge against matches against sites with many ambiguous characters. Such sites would score well in terms of percentage mismatch, but have a poor log-likelihood ratio.

The log-likelihood ratio for strings is computed roughly as follows. Each match against an

unambiguous base is worth a LLR of 2. A match against a ambiguity code that represents two bases, e.g., S=C or G, is worth 1. A match against an ambguity code that represents three bases, e.g., D = A, G, or T is worth about 0.75. A match against an 'N' is worth 0.

The parameter's value is constrained to be >=0 .

The parameter's type is real number. The standard decimal format plus scientific notation (eg, 2.1e-32.1). The decimal point is optional. An empty string defaults to '0.0'. Use commas or white space to make the number more legible.

Minimum string length (t_w):

TESS will not search for sites that are shorter than this length.

Set this value to higher values to avoid getting swamped with weak hits.

The parameter's type is integer. A series of digits optionally preceeded by a minus sign. Commas are ok.

Secondary Lg-Likelihood Deficit:

This threshold is used to highlight alignments that are especially good. Alignments that fail to meet this threshold are reported but are indicated in the sequence display by magenta or cyan rather than red or blue.

See also: mlld

The parameter's value is constrained to be >=0.0 and <=6.0.

The parameter's type is real number. The standard decimal format plus scientific notation (eg, 2.1e-32.1). The decimal point is optional. An empty string defaults to '0.0'. Use commas or white space to make the number more legible.

Count significance threshold:

This threshold is used to remove those matrices that do not produce a significantly high number of hits in the sequence. The total number of hits is tallied for PWM. The number of hits is approximated by a Poisson distribution with a rate estimated from empirical data measured on a random sequence generated from a uniform distribution using the log-likelihood or similarity threshold. A p-value (the probability of getting the same or more hits) is computed for each matrix. If you select this option then those PWMs with a p-value greater (more likely) than the threshold are eliminated.

The parameter's value is constrained to be >=0.0 and <=1.0.

The parameter's type is real number. The standard decimal format plus scientific notation (eg, 2.1e-32.1). The decimal point is optional. An empty string defaults to '0.0'. Use commas or white space to make the number more legible.

Am. J. Hum. Genet. 67:788–792, 2000

INVITED EDITORIAL
Genomic Sequence, Splicing, and Gene Annotation

Stephen M. Mount

Department of Cell Biology and Molecular Genetics, University of Maryland, College Park

Introduction

The sequence of the human genome is at hand. Most scientists who use the sequence will rely on annotations that provide information about the number and location of genes and about their inferred protein products. Traditionally, genes have been annotated by scientists with a particular interest in them. However, annotation of the complete human genome sequence will have to be at least partially automated. Gene annotation incorporates cDNA data (including expressed sequence tags [ESTs]), sequence similarity, and computational predictions based on the recognition of probable splice sites and coding regions (Stormo 2000; also see David Haussler's Web site, Computational Genefinding). The state of the art was recently surveyed by the Genome Annotation Assessment Project-GASP1 and must be regarded as imperfect (Bork 2000; Reese et al. 2000).

This review enumerates aspects of pre-mRNA splicing that limit our ability to predict gene structure from genomic sequence, drawing on the recently annotated complete genome of *Drosophila melanogaster* (Adams et al. 2000) as an example. In particular, the following four facts will be discussed. First, splice sites do not always conform to consensus. Second, noncoding exons are common. Third, internal exons can be arbitrarily small, and small internal exons confound not only gene finding but also the alignment of cDNA and genomic sequences. Fourth, splice sites are not recognized in isolation, and nucleotides that are far from splice sites can affect splicing. This list and the accompanying analysis should make molecular geneticists aware of the ways in which gene annotations can be wrong and should encourage recourse to the primary data. In addition, the same considerations indicate that inherited disease can be caused by mutations remote from splice sites that nevertheless affect splicing.

Received August 3, 2000; accepted for publication August 15, 2000; electronically published September 8, 2000.

Address for correspondence and reprints: Dr. Stephen M. Mount, Department of Cell Biology and Molecular Genetics, H. J. Patterson Hall, University of Maryland, College Park, MD 20742-5815. E-mail: sm193@umail.umd.edu

This article represents the opinion of the author and has not been peer reviewed.

0002-9297/2000/6704-0003$02.00

Discussion

Splice Sites Do Not Always Conform to Consensus

It is well established that nearly all splice sites conform to consensus sequences (Mount 1982; Senapathy et al. 1990; Zhang 1998). These consensus sequences include nearly invariant dinucleotides at each end of the intron—GT at the 5′ end of the intron and AG at the 3′ end of the intron. Most gene-finding software and most human annotators will find only introns that begin with a GT and end with an AG. However, nonconsensus splice sites have been described, and I will discuss three classes, in decreasing order of frequency.

The most common class of nonconsensus splice sites consists of 5′ splice sites with a GC dinucleotide. Senapathy et al. (1990) listed 17 examples among 3,724 5′ splice sites, suggesting a frequency of ~0.5%. Jackson (1991) listed a total of 26 GC sites, whereas Wu and Krainer (1999) cited an additional 18 examples. GC 5′ splice sites are consistent with the experimental observation that, of the six possible point mutations within the GT dinucleotide, mutation of T to C in position 2 has the smallest effect on in vitro splicing (Aebi et al. 1986). At other positions within the consensus, GC sites conform extremely well to the standard consensus; for example, 42 of the 44 sites cited above have a consensus G residue at both position −1 and position +5. It is reasonable to assume that GC sites are recognized by the standard (U2-dependent) spliceosome.

The second class of exception to splice-site consensus is U12 introns, a minor class of rare introns with splice-site sequences that are very different from the standard consensus but that are very similar to each other. The existence of this class was first pointed out by Jackson (1991) and was considered in more detail by Hall and Padgett (1994). It was subsequently discovered that U12 introns are removed by a minor spliceosome containing the rare U11, U12, U4atac, and U6atac snRNPs, in place of U1, U2, U4, and U6 (Tarn and Steitz 1997; Burge et al. 1998). Some U12 introns have AT and AC in place of GT and AG and are known as "AT-AC" introns. However, terminal intron dinucleotide sequences do not

distinguish between U2- and U12-dependent introns (Dietrich et al. 1997). Rather, U12 introns can be identified by highly conserved sequences at the 5′ splice site (RTATCCTY; R = A or G, and Y = C or T) and branch site (TCCTRAY). U12 introns are found in many eukaryotes, including *Drosophila melanogaster* (Adams et al. 2000) and *Arabidopsis thaliana* (Shukla and Padgett 1999) but not *Caenorhabditis elegans*.

Finally, there are a small number of nonconsensus sites that fit into neither of the two categories mentioned above. Many reports of such variant splice sites can be traced to errors in annotation or interpretation, polymorphic differences between the sources of cDNA and genomic sequence, inclusion of pseudogene sequences, or failure to account for somatic mutation (author's unpublished data; for examples, see Jackson 1991). However, there are many examples of sites that match the consensus very poorly, and experimental work has established that 5′ splice sites do not absolutely require GT—and that 3′ splice sites do not absolutely require AG—in order to be recognized in vivo (Aebi et al. 1986; Roller et al. 2000, and references therein). In yeast, an intron that is within the HAC1 mRNA and that has no similarity to the standard nuclear pre-mRNA intron consensus sequence is spliced by a specific, regulated, endonuclease and tRNA ligase (Sidrauski et al. 1996). This intron provides a precedent for introns in protein-coding genes with completely novel splice sites.

Noncoding Exons Are Common

There is considerable confusion between exons and coding regions. The term "exon" was coined by Gilbert (1978) to refer to what is left when introns are removed by splicing, and RNAs that are entirely noncoding (such as tRNAs) are sometimes spliced. However, the term exon is often misused to refer to a stretch of coding information. In reality, however, noncoding exons are quite common, occurring in >35% of human genes (Zhang 1998). Gene-finding software generally detects sequence features characteristic of coding regions rather than of exons and does not even attempt to identify noncoding exons, or noncoding portions of exons. This is because the statistical biases introduced by protein-coding are in fact a very powerful tool for the identification of coding DNA, and no similar tool has been developed for the identification of noncoding exons.

A similar problem can arise in genes without noncoding exons. If the first intron occurs near the initiator AUG, then the coding information in the first exon can be very short and difficult to identify by measures of coding tendency. Furthermore, the first intron tends to be longer than average (Maroni 1996), and such an arrangement can separate promoter function (perhaps including downstream transcriptional enhancer elements lying in the first intron) from the bulk of the coding information downstream. In these cases, investigators have no way of knowing how much information is missing—or where the 5′ end of the gene is likely to reside—without experimental data such as a cDNA sequence or a 5′ EST.

Internal Exons Can Be Arbitrarily Small

A less frequent but perhaps more serious problem for gene-discovery methods is posed by small internal exons. Vertebrate internal exons have an average size of ~130 nucleotides (Hawkins 1988; Zhang 1998), and roughly 65% of internal human exons are 68–208 nucleotides in length (Maroni 1996). This size distribution reflects a functional constraint. Optimal splicing efficiency requires exons with sizes of ~50–300 nucleotides (Robberson et al. 1990; Dominski and Kole 1991; see review by Berget 1995). However, a considerable number, >10%, of exons are <60 nucleotides in length, and it is these exons that can be difficult to identify by measures of coding tendency.

Just how small can internal exons be? There appears to be no lower limit, and many cases of exons <10 nucleotides have been described (for examples, see Stamm et al. 1994; also see the author's Web site, Gene Annotation and Splice Site Selection). An illustrative case is the *invected* gene of *D. melanogaster* (also listed in GadFly as CG17835). This gene encodes a homeodomain protein that is similar to *engrailed*, and these two genes are adjacent. One of four *invected* exons is only 6 nucleotides long and is flanked by introns of 27,659 and 1,134 nucleotides. Significantly, this exon is not recognized by cDNA alignment software such as SIM4 (Florea et al. 1998), and the gene is incorrectly annotated (GenBank accession number AE003825.1). As a result, the protein sequence predicted by annotation of the genome (Adams et al. 2000; GenBank accession number AAF58640) differs from that predicted from the cDNA (Coleman et al. 1987; GenBank accession number CAA28885), because of a frameshift affecting the entire carboxyl-terminal coding exon, a highly conserved region of the protein. This is despite the fact that the microexon sequence, GTCGAA, is flanked by intron sequences that perfectly match the splice-site consensus. Use of this microexon provides perfect agreement between the cDNA and genomic sequences when consensus splice sites are used, whereas the annotation predicts an RNA with several discrepancies relative to the cDNA. The frameshift is due to the predicted use of a 5′ splice site 10 nucleotides downstream of the true 5′ splice site, which was apparently selected to account for the microexon. It seems clear that the protein sequence predicted by the cDNA is correct. Why was it not incorporated into the annotation? The alignment problem

arises because a pattern-matching algorithm that locates exons by similarity between the cDNA and the genomic sequence cannot find exons of this size unless its stringency is reduced to an unacceptable level (Florea et al. 1998).

The notion that exons can be arbitrarily small is supported by the observation of exons with length 0. Of course, such sites are not exons at all but, rather, are resplicing sites (see fig. 1). This phenomenon has been demonstrated in the case of the *Drosophila Ultrabithorax* locus (Hatton et al. 1998), which has a region of 60 kb containing two alternatively spliced exons, and may be a general feature of long introns (J. Burnette and A. J. Lopez, personal communication). The existence of resplicing sites not only illustrates the lack of a lower limit to exon size (which has implications for gene annotation) but also has implications for the analysis of hereditary mutations. A mutation at one of these sites could potentially create a frozen intermediate such as that diagrammed in figure 1. This partially spliced RNA would probably be unstable, because of nonsense-mediated decay (Culbertson 1999), and the apparent result would be no RNA (rather than aberrantly spliced RNA). Such mutations would be very hard to identify.

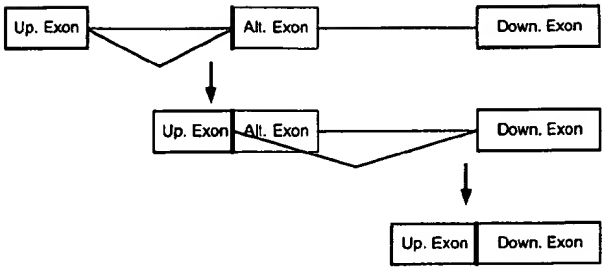

Figure 1 Small internal exons and resplicing. This schematic figure indicates the pathway of resplicing demonstrated for the *Drosophila Ubx* locus (Hatton et al. 1998). The thicker vertical line indicates a resplicing site, which does not contribute any nucleotides to the final mRNA product. The same pathway could be followed in the case of a microexon, in which case an arbitrarily small number of nucleotides would remain in the mRNA product. "Up. Exon" and "Down. Exon" denote the exons upstream and downstream of the resplicing site, respectively. In the case of *Ubx*, the sequence immediately downstream of the resplicing site is an alternatively spliced exon (here designated "Alt. Exon"), but resplicing sites are not always accompanied by such alternatively spliced exons (J. Burnette and A. J. Lopez, personal communication).

Nucleotides Far from Splice Sites Can Affect Splicing

No method of evaluating potential splice sites that is based on sequence alone can be 100% reliable. One can be sure of this because many sequences that are not splice sites are capable of acting as splice sites, and vice versa. Perhaps the clearest demonstration of this is provided by the activation of cryptic splice sites. These are splice sites that are used, sometimes with 100% efficiency, when a natural splice site has been mutationally inactivated. The activation of cryptic sites occurs in approximately one-third of splicing mutations (Nakai and Sakamoto 1994). The phenomenon shows that the cryptic sites are perfectly capable of being recognized by the splicing machinery. Clearly, the sequence of such cryptic sites is compatible with splicing, and context is important for splice-site choice.

Two contextual elements that contribute to splice-site selection are the location of splice sites relative to each other and splicing-enhancer sequences. The exon-size preferences described above are widely understood in terms of an exon-definition model that includes the interaction of splicing factors bound at either end of an exon (Berget 1995). The requirement for productive interactions among splicing factors, including U1 snRNPs at the 5′ splice site and U2 snRNP auxiliary factor (U2AF) at the 3′ splice site, are thought to give rise to preferred exon lengths because of steric constraints and geometry favoring interactions. In the case of small introns, a similar model of intron bridging has been proposed (Guo and Mount 1995; McCullough and Berget 1997). In combination, these models suggest that, in order to be recognized, a splice site must have a partner an appropriate distance away, so that either exon definition or intron definition is facilitated by the spacing. One experimental distinction between exon definition and intron definition is the result of mutations that inactivate the splice site. Failure to undergo exon definition results in exon skipping, whereas failure to undergo intron definition results in intron retention.

Not only is the use of one splice site dependent on the presence of its partner across the exon, but weakness in one partner can be compensated by strength in the other, as seen with second-site revertants of splice-site mutations that cause exon skipping. In an analysis of splicing mutations at the dihydrofolate reductase locus, Carothers et al. (1993) found that a mutation at the 5′ splice site of exon 5 (G to C in the third position of the intron) could be partially reversed by mutations that increased the strength of the 3′ splice site upstream of the same exon (AAAG| to TTAG|, ACAG|, or ATAG|). Although reversion was not complete, these data provide a strong argument that whether a sequence functions as a splice site depends not only on its intrinsic strength but also on its context. Similarly, there are mutations that create splice sites within introns, activating cryptic exons by recruitment of appropriately placed partners (e.g., see Bagnall et al. 1999).

Splicing enhancers are sequences that stimulate splicing at nearby sites. A family of non-snRNP splicing factors known as "SR proteins" appear to be important for the recognition of splicing enhancers in

exons (Blencowe 2000). A splicing difference between SMN1 and SMN2, which explains their differential effects on spinal muscular atrophy, has been attributed to a translationally silent substitution within the coding sequence that affects splicing (Lorson et al. 1999). Similarly, H.-X. Liu, L. Cartegni, M. Q. Zhang, and A. R. Krainer (personal communication) have shown that a nonsense mutation causing the skipping of BRCA1 exon 18 affects splicing in vitro and that a missense mutation at the same position can also cause exon skipping. There are also splicing-enhancer sequences in introns—and examples of mutations that affect them (Cogan et al. 1997). Although general mechanisms for their function have yet to be defined, there is some evidence that at least some splicing enhancers in introns may act by facilitating exon definition in the case of small exons (Carlo et al. 2000).

Outlook

This review has presented aspects of pre-mRNA splicing that pose special problems for gene annotation. However, even though the best gene finders predict genes exactly right less than half the time, 95% of total coding nucleotides are predicted accurately, and <5% of genes are completely missed (Reese et al. 2000; Genome Annotation Assessment Project-GASP1). When cDNA and homology data are available, annotations will tend to be even better. Thus, one would be wrong to conclude from this review that the gene annotations attending the human genome sequence will not provide an extremely valuable resource. Nevertheless, molecular geneticists will want to have an understanding of the kinds of errors that are likely to occur—and to carefully review the available evidence for genes that matter to them. Annotators are likewise obligated to make the source of each specific aspect of their annotation an integral part of the annotation; for example, if part of the annotation is supported by a EST whereas the rest of it is based on the prediction of a gene finder, then the limits of the cDNA should be indicated, and the accession number of the EST should be part of the annotation.

A related but distinct point is that these same factors are also relevant when candidate mutations are evaluated during the analysis of hereditary disease. Mutations that lie within splicing enhancers, at resplicing sites, or at cryptic splice sites can affect splicing even when they lie some distance from the splice sites actually used in the generation of the affected mRNA. The problem is further compounded by alternative splicing and the interplay between splicing and polyadenylation, topics that are beyond the scope of the present review.

In summary, gene annotations will be a valuable resource. However, they will not substitute for expertise in molecular genetics.

Acknowledgments

Support by National Institutes of Health grant GM37991-11 is gratefully acknowledged. I thank Doug Black for helpful comments on the manuscript. I thank James Burnette, A. Javier Lopez, and Adrian Krainer for providing information prior to publication.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Computational Genefinding, http://www.cse.ucsc.edu/~haussler/genefindingpaper

GadFly: Genome Annotation Database of Drosophila, http://www.fruitfly.org/annot/index.html

GenBank, http://www.ncbi.nlm.nih.gov/ (for incorrect annotation of *invected* [accession number AE003825.1] and predicted protein sequence [accession numbers AAF58640 and CAA28885])

Gene Annotation and Splice Site Selection, http://www.wam.umd.edu/~smount/Annotation.html

Genome Annotation Assessment Project-GASP1, http://www.fruitfly.org/GASP1/index.html

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, et al (2000) The genome sequence of Drosophila melanogaster. Science 287:2185–2195

Aebi M, Hornig H, Padgett RA, Reiser J, Weissman C (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. Cell 47:555–565

Bagnall RD, Waseem NH, Green PM, Colvin B, Lee C, Giannelli F (1999) Creation of a novel donor splice site in intron 1 of the factor VIII gene leads to activation of a 191 bp cryptic exon in two haemophilia A patients. Br J Haematol 107:766–771

Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem 270:2411–2414

Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 25:106–110

Bork P (2000) Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res 10:398–400

Burge CB, Padgett RA, Sharp PA (1998) Evolutionary fates and origins of U12-type introns. Mol Cell 2:773–785

Carlo T, Sierra R, Berget SM (2000) A 5′ splice site-proximal enhancer binds SF1 and activates exon bridging. Mol Cell Biol 20:3988–3995

Carothers AM, Urlaub G, Grunberger D, Chasin LA (1993) Splicing mutants and their second-site suppressors at the dihydrofolate reductase locus in Chinese hamster ovary cells. Mol Cell Biol 13:5085–5098

Cogan JD, Prince MA, Lekhakula S, Bundrey S, Futrakul A, McCarthy EM, Phillips JA III (1997) A novel mechanism of aberrant pre-mRNA splicing in humans. Hum Mol Genet 6:909–912

Coleman KG, Poole SJ, Weir MP, Soeller WC, Kornberg T

(1987) The invected gene of Drosophila: sequence analysis and expression studies reveal a close kinship to the engrailed gene. Genes Dev 1:19–28

Culbertson MR (1999) RNA surveillance: unforeseen consequences for gene expression, inherited genetic disorders and cancer. Trends Genet 15:74–80

Dietrich RC, Incorvaia R, Padgett RA (1997) Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol Cell 1:151–160

Dominski Z, Kole R (1991) Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol 11:6075–6083

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA. Genome Res 8:967–974

Gilbert W (1978) Why genes in pieces? Nature 271:501

Guo M, Mount SM (1995) Localization of sequences required for size-specific splicing of a small *Drosophila* intron. J Mol Biol 253:426–437

Hall SL, Padgett RA (1994) Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. J Mol Biol 239:357–365

Hatton AR, Subramaniam V, Lopez AJ (1998) Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. Mol Cell 2:787–796

Hawkins JD (1988) A survey on intron and exon lengths. Nucleic Acids Res 16:9893–9905

Jackson IJ (1991) A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res 19:3795–3798

Lorson CL, Hahnen E, Androphy EJ, Wirth B (1999) A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Proc Natl Acad Sci USA 96:6307–6311

Maroni G (1996) The organization of eukaryotic genes. Evol Biol 29:1–19

McCullough AJ, Berget SM (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol Cell Biol 17:4562–4571

Mount SM (1982) A catalogue of splice junction sequences. Nucleic Acids Res 10:459–472

Nakai K, Sakamoto H (1994) Construction of a novel database containing aberrant splicing mutations of mammalian gene. Gene 141:171–177

Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE (2000) Genome annotation assessment in Drosophila melanogaster. Genome Res 10:483–501

Robberson BL, Cote GL, Berget SM (1990) Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol 10:84–94

Roller AB, Hoffman DC, Zahler AM (2000) The allele-specific suppressor sup-39 alters use of cryptic splice sites in Caenorhabditis elegans. Genetics 154:1169–1179

Senapathy P, Sharpiro MB, Harris NL (1990) Splice junctions, branch point sites and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol 183:252–278

Shukla GC, Padgett RA (1999) Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. RNA 5:525–538

Sidrauski C, Cox JS, Walter P (1996) tRNA ligase is required for regulated mRNA splicing in the unfolded protein response. Cell 87:405–413

Stamm S, Zhang MQ, Marr TG, Helfman DM (1994) A sequence compilation and comparison of exons that are alternatively spliced in neurons. Nucleic Acids Res 22:1515–1526

Stormo GD (2000) Gene-finding approaches for eukaryotes. Genome Res 10:394–397

Tarn WY, Steitz JA (1997) Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. Trends Biochem Sci 22:132–137

Wu A, Krainer AR (1999) AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. Mol Cell Biol 19:3225–3236

Zhang MQ (1998) Statistical features of human exons and their flanking regions. Hum Mol Genet 7:919–932

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Jan. 1998, p. 82–87
0099-2240/98/$04.00+0

Vol. 64, No. 1

The Sequence of Spacers between the Consensus Sequences Modulates the Strength of Prokaryotic Promoters

PETER RUHDAL JENSEN* AND KARIN HAMMER

Department of Microbiology, Technical University of Denmark, DK-2800 Lyngby, Denmark

Received 7 July 1997/Accepted 21 October 1997

We constructed a library of synthetic promoters for *Lactococcus lactis* in which the known consensus sequences were kept constant while the sequences of the separating spacers were randomized. The library consists of 38 promoters which differ in strength from 0.3 up to more than 2,000 relative units, the latter among the strongest promoters known for this organism. The ranking of the promoter activities was somewhat different when assayed in *Escherichia coli*, but the promoters are efficient for modulating gene expression in this bacterium as well. DNA sequencing revealed that the weaker promoters (which had activities below 5 relative units) all had changes either in the consensus sequences or in the length of the spacer between the −35 and −10 sequences. The promoters in which those features were conserved had activities from 5 to 2,050 U, which shows that by randomizing the spacers, at least a 400-fold change in activity can be obtained. Interestingly, the entire range of promoter activities is covered in small steps of activity increase, which makes these promoters very suitable for quantitative physiological studies and for fine-tuning of gene expression in industrial bioreactors and cell factories.

Metabolic engineering has promising perspectives with respect to improving the properties and performances of microorganisms used as industrial bioreactors, as cell factories, and in food fermentations. The importance of tuning gene expression in this context, i.e., to perform metabolic optimization rather than massive overexpression or gene inactivation, is now far more appreciated. However, the more subtle approach of metabolic optimization is hampered by the lack of proper expression systems for tuning gene expression in many microorganisms. Also, the fundamental understanding of a biological system through metabolic control analysis (5, 10) requires the tuning of enzyme activities in order to calculate the so-called control coefficients. For some organisms, expression systems that allow for changing gene expression for scientific purposes and for a limited set of experimental conditions have been developed. Thus, for *Escherichia coli*, the *lac* system, the *c*I-regulated lambda p_R/p_L, and many derivatives of these systems have been widely applied, and such systems have also been adapted for use in other organisms (for a recent review, see reference 12). With respect to changing steady-state gene expression, these systems can sometimes be difficult to apply, particularly when it comes to changing gene expression on an industrial scale. Besides, in most food fermentation processes, the addition of chemicals as inducers of gene expression or the changing of other process parameters is not acceptable; in such cases, there are virtually no expression systems available for tuning gene expression and thus for performing accurate metabolic optimization.

Lactic acid bacteria are widely used in food fermentation, e.g., cheese and yoghurt production, but besides lactic acid, these bacteria excrete a spectrum of organic compounds. Some of these are desirable with respect to the development of texture and flavors or for bioconservation purposes, and some are undesirable for similar or different reasons. The lactic acid bacteria are therefore obvious candidates for attempts to optimize the pattern of formation of these compounds for specific applications. But the experimental tools for manipulating gene expression are not well developed for these bacteria. An exception is the nisin-inducible system, developed recently by de Ruyter et al. (2). This system appears to be well suited for inducing gene expression in *Lactococcus lactis* by adding the antibiotic nisin (which is accepted as a food additive). A question that perhaps needs to be addressed in this context is whether the nisin expression system is also suitable for achieving a steady level of gene expression. In addition, for effective metabolic optimization, it is often necessary to optimize the expression of a number of genes, which is not feasible with the systems developed so far.

Here we describe a method for tuning steady-state gene expression in *L. lactis*. We overcome many of the limitations discussed above by using libraries of synthetic promoters which cover a wide range of promoter activities and show that the strength of prokaryotic promoters can be modulated by randomizing the spacer sequences that separates the consensus sequences. The system is food grade and well suited for use in industrial bioreactors and food fermentation processes. In addition, the system should be applicable to a broad range of biological systems. (Potential commercial users should be aware that the approach for obtaining the synthetic promoters, as well as the promoter sequences, were filed for patent worldwide [7a]).

MATERIALS AND METHODS

Bacterial strains and plasmids. The *E. coli* K-12 strain BOE270 (1) is highly competent with respect to transformation and was derived from strain MT102, which in turn is an *hsdR* derivative of strain MC1000 [*araD139* Δ(*ara-leu*)*7679 galU galK* Δ(*lac*)*174 rpsL thi-1* (1a))]. BOE270 was used for studying promoter activities in *E. coli* as well as for cloning purposes and propagation of plasmid DNA in *E. coli*. The plasmid-free *L. lactis* subsp. *cremoris* strain MG1363, which does not express β-galactosidase activity (4), was used for studying promoter activities in *L. lactis*.

The promoter cloning vector pAK80 (7) was used for cloning the synthetic promoters DNA fragments. pAK80 is a shuttle vector for *L. lactis* and *E. coli*, conferring erythromycin resistance to the host cells. The vector carries the promoterless *lacL* and *lacM* genes from *Leuconostoc lactis* (which codes for β-galactosidase enzyme activity). It contains a multiple cloning site for the insertion of DNA fragments harboring putative promoter signals, just upstream

* Corresponding author. Mailing address: Department of Microbiology, Technical University of Denmark, Building 301, DK-2800 Lyngby, Denmark. Phone: 45 45252510. Fax: 45 45932809. E-mail: prj@im.dtu.dk.

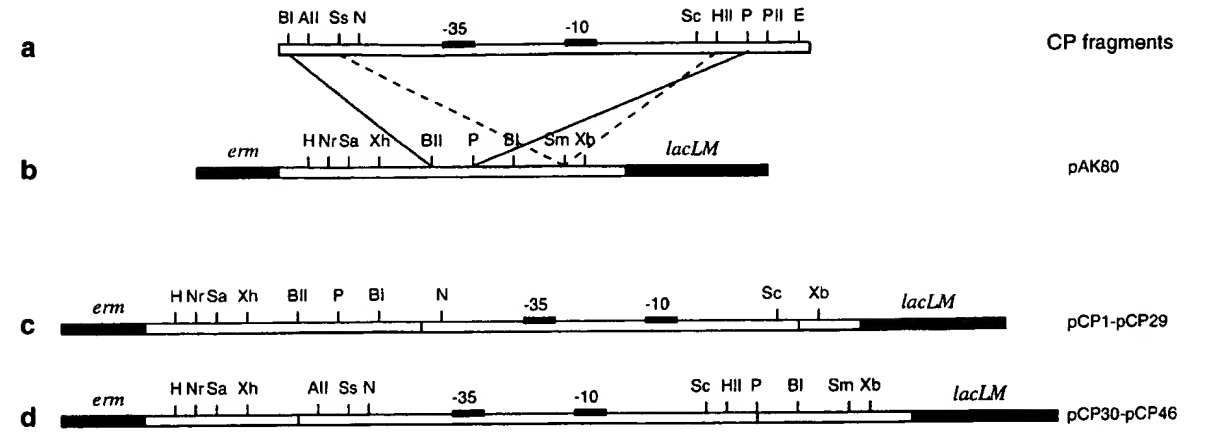

FIG. 1. Strategies used for cloning synthetic promoter fragments into the promoter cloning vector pAK80. (a) Double-stranded DNA fragments carrying putative promoter activities. (b) Restriction map and schematic representation of the relevant parts of the promoter cloning vector. The stippled and solid lines show the strategies used for cloning pCP1 through pCP29 and pCP30 through pCP46, respectively. (c) Restriction map of clones pCP1 through pCP29. (d) Restriction map of clones pCP30 through pCP46. Note that a number of clones have been subject to cloning artifacts and thus may have a slightly different restriction map. BI, *Bam*HI; AII, *Afl*II; Ss, *Ssp*I; N, *Nsi*I (*Pst*I compatible); Nr, *Nru*I; Sc, *Sca*I; HII, *Hin*cII; P, *Pst*I; PII, *Pvu*II; E, *Eco*RI; Sa, *Sac*I; Xh, *Xho*I; BII, *Bgl*II; Sm, *Sma*I; Xb, *Xba*I (not drawn to scale).

the promoterless *lacL* and *lacM* genes from *Leuconostoc lactis*. Together, the *lacL* and *lacM* genes codes for a β-galactosidase.

Enzymes. Restriction enzymes, Klenow DNA polymerase, calf intestine phosphatase, and T4 DNA ligase were obtained from and used as recommended by Pharmacia and New England Biolabs.

Oligonucleotides. Oligonucleotides were obtained from Hobolth DNA Synthesis (Hillerød, Denmark).

Second-DNA-strand synthesis. The single-stranded promoter oligonucleotides were converted to double-stranded DNA, using a 10-bp oligonucleotide (5′-CC GAATTCAG) complementary to the 3′ end of the promoter oligonucleotide as primer for the second-strand synthesis by the Klenow fragment of DNA polymerase I.

Cloning of synthetic DNA fragments into the promoter cloning vector pAK80. Two different cloning strategies were used (Fig. 1). In strategy A, the mixture of DNA fragments was digested with two restriction enzymes, *Hin*cII and *Ssp*I, and pAK80 was digested with *Sma*I. In strategy B, the mixture of DNA fragments was digested with two restriction enzymes, *Bam*HI and *Pst*I, and pAK80 was digested with *Bgl*II and *Pst*I. In both strategies, the promoter fragments were then ligated to the compatible vector fragments. The ligation mixtures were then transformed into Ca^{2+}-competent cells (13) by using a standard transformation procedure (13), and the transformation mixture were plated (at 30°C) on LB plates containing erythromycin (200 μg/ml) and 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal; 100 μg/ml). A total of 150 erythromycin-resistant transformants were obtained; all were white initially, but after prolonged incubation (up to 2 weeks at 4°C), a number had become blue to various extents. Later, we discovered that the development of blue color from *E. coli* colonies (but not *L. lactis* colonies) expressing *lacLM* is greatly enhanced by adding 1% glycerol to the transformation plates (data not shown). Plasmids were isolated from these blue colonies, and it was confirmed by restriction enzyme analysis that most of these clones had promoter fragments inserted in the multiple cloning site of pAK80, in the orientation that would direct transcription into the β-galactosidase gene (*lacLM*). The 46 colonies isolated had become blue to various extents; 29 from cloning strategy A (containing plasmids pCP1 through pCP29) and 17 from strategy B (containing plasmids pCP30 through pCP46) were picked for further analysis. The two weakest promoter clones, pCP31 and pCP43, did not contain a promoter fragment, and four promoter clones, pCP18, pCP19, pCP33, and pCP44, turned out to be identical to pCP27, pCP22, pCP35, and pCP45, respectively. Indeed, the activities of these sets were almost identical, which also demonstrates the reproducibility of the assay used here. The chances that two identical sequences would have arisen by coincidence during the oligonucleotide synthesis is of course negligible, and these four clones must therefore be the result of a cell division that took place after the plasmids were transformed but before the cells were plated.

Transformation of *L. lactis*. Cells of *L. lactis* subsp. *cremoris* MG1363 (4) were made competent by growth overnight in GM17 medium containing 2% glycine as described by Holo and Ness (6). Plasmid DNA from the 46 clones described above was then transformed into these cells by electroporation (6). The cells were allowed to regenerate in SGM17 medium for 2 h and then plated on SR plates containing erythromycin (2 μg/ml) and X-Gal (100 μg/ml).

β-Galactosidase assay. The assay was done as described by Miller (14) and modified by Israelsen et al. (7). Cultures carrying the plasmid derivatives of pAK80 were grown in rich medium overnight at 30°C. The medium used for *L. lactis* was M17 medium supplemented with erythromycin (2 μg/ml) and 1% glucose; for *E. coli*, LB medium supplemented with erythromycin (200 μg/ml) was used. The results presented are averages of measurements of the activities of at least three individual cultures of each clone. The standard errors were less than 30% for *E. coli* activities and less than 20% for *L. lactis* activities. Aliquots of 25 to 100 μl of the cultures were used in the β-galactosidase assay except in the case of the weakest promoter clones, where up to 2 ml of culture was concentrated and used in the assay.

RESULTS

The purpose of this work was to generate a library of synthetic constitutive promoters as a tool for genetic engineering of *L. lactis*. The promoters should cover a wide range of promoter activities, in small steps of activity changes, so that they would be applicable to quantitative physiological studies and for metabolic optimization. The following strategy was used: (i) design and synthesize a degenerated oligonucleotide sequence that encodes consensus sequences for *L. lactis* promoters, separated by spacers of random sequences; (ii) convert this mixture of oligonucleotides to double-stranded DNA fragments, using DNA polymerase and a short oligonucleotide primer complementary to the 3′ end of the degenerated oligonucleotide; and (iii) clone this mixture of DNA fragments into a promoter probing vector. The idea behind this strategy is that even though the consensus sequences should be important elements of an efficient promoter, the context in which the consensus sequences are located may modulate the strength of the promoters to some extent.

Design and construction of synthetic promoters for *L. lactis*. A considerable number of promoters have been cloned and sequenced from *L. lactis* (see the review by de Vos and Simons [3]). From these data, we extracted extended consensus sequence motifs for *L. lactis* promoters (Fig. 2A). The Pribnow box or the −10 sequence TATAAT and the −35 sequence TTGACA, known to be present in many prokaryotic promoters, are also well conserved for *L. lactis*. In addition, the sequence TG is often found 1 bp upstream of the −10 sequence; it is also possible to determine a consensus sequence for the 4 bp immediately upstream of the −35 motif, ATTC. Nilsson and Johansen (16) found well-conserved sequences among promoters of the rRNA operons: AGTTT at position −44 and GTACTGTT at positions +1 to +8. In addition to these mo-

A.

```
5' CATNNNNNAGTTTATTCTTGACANNNNNNNNNNNNNNNTGRTATAATANNWNAGTACTGTT 3'
          -44      -35                    -15  -10        +1
```

B.

```
          AflII
    BamHI       SspI  NsiI
  1
5' CGGGATCCTTAAGAATATTATGCATNNNNNAGTTTATTCTTGACANNNNNNNNNNNNNT
                                  -44     -35                -15

                          PstI    EcoRI
                 ScaI  HincII  PvuII

GRTATAATANNWNAGTACTGTTAACTGCAGCTGAATTCGG 3'
  -10       +1
```

FIG. 2. Oligonucleotide sequence used for the generation of a library of synthetic promoters for *L. lactis*. (A) Consensus sequence for *L. lactis* promoters derived from data published in the literature. N = 25% each A, C, G, and T; R = 50% each A and G; W = 50% each A and T. (B) The design of the oligonucleotide. The sequence contains a number of recognition sequences for restriction endonucleases, for use in the subsequent cloning strategy. Note that the sequence from positions +1 to +8, which is a putative stringent response site, can be deleted in the cloning process if necessary. See text for further details.

tifs, two semiconserved base pairs were included, R (=A or G) upstream of the −10 sequence and W (=A or T) at position −3. Based on these data, we designed an oligonucleotide which also encodes recognition sites for multiple restriction enzymes (Fig. 2B). This mixture of oligonucleotides was converted to double-stranded DNA fragments, using a short primer complementary to the 3′ end. Finally, the resulting double-stranded DNA fragments, encoding potential promoter structures, were cloned into the polylinker on the promoter probe vector, pAK80 (7), upstream of the promoterless β-galactosidase gene, using *E. coli* as a host; this resulted in plasmids pCP1 through pCP46.

Activities of the synthetic promoters in *L. lactis*. Plasmids, pCP1 through pCP46 were then transformed into *L. lactis* subsp. *cremoris* MG1363. The different plasmids gave rise to colonies exhibiting very different intensities of blue on plates containing X-Gal. The specific activities of β-galactosidase in liquid cultures of these clones were then determined (Fig. 3) and found to vary from 0.3 Miller unit, or from slightly above the activity found with the cloning vector pAK80 without any insert, to up to more than 2,000 Miller units. Together, the promoters covered 3 to 4 logs of promoter activities in small steps of activity change.

Sequence analysis of the CP promoters. A very interesting point is the molecular basis for the differences in strength of the CP promoters, and we therefore took on the task of sequencing the promoter clones. Eighteen clones were perfect in the sense that they had the DNA sequence that was specified by the oligonucleotide (Fig. 4). The activities of these 18 promoter clones covered, in small steps of activity change, a 50-fold range of activity, from 34 up to 1,800 Miller units. Four of the CP promoters had a 16-bp spacer between the −35 and −10 sequences instead of the 17 bp specified in the oligonucleotide sequence, and the activities carried by these four clones were weak, ranging from 0.7 to 12 Miller units. Four clones had base pair changes in the −35 sequence, and two had base pair changes in the −10 sequence; those clones also had rather weak activity (0.3 to 69 Miller units).

Some clones had 1-bp deletions or a base pair change outside the −35 to −10 region or have been subject to other cloning artifacts. However, the activities of these promoter clones were all within the range covered by the perfect clones, i.e., activities from 58 to 2050 Miller units, which indicates that in this case, consensus sequences outside the −35 to −10 sequence are of little importance with respect to determining the promoter strength.

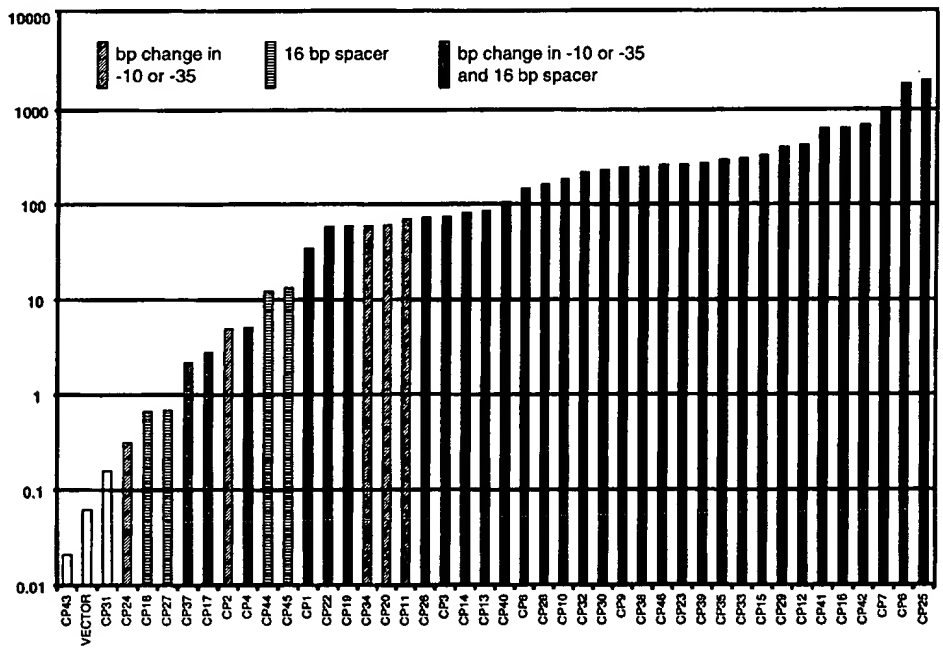

FIG. 3. Library of synthetic promoters for *L. lactis*. Promoter activities (Miller units) were assayed from the expression of a reporter gene (*lacLM*) encoding β-galactosidase transcribed from the different synthetic promoter clones on the promoter cloning vector pAK80. The patterns of the data points indicate which promoter clones contain errors in either the −35 or the −10 consensus sequence or in the length of the spacer between these sequences.

Promoter		Spec. β-gal activity	
		L. lactis	*E. coli*
Oligoseq	CATNNNNNAGTTTATTCTTGACANNNNNNNNNNNNNNNTGRTATAATANNWNAGTACTGTT		
CP24	CATGGGTAAGTTTATTCTTCACACTATCTGGGCCCGATGGTATAATAAGTGACTACTGTT	0.3	3.1
CP18=CP27	CATTTTGCAGTTTATTCTTGACATTGTGTGCTTCGGGTG-TATAATA-CTAAGTACTGTT	0.7	0.2
CP37	CATCATTAAGTTTATTCTTCACATTGGCCGGAATTGTTG-TATAATACCTTAGTACTGTT	2.2	16
CP17	CATTCTGGAGTTTATTCTTGAC-CGCTCAGTATGCAGTGGTATAATAGTACAGTACTGTT	2.7	10
CP2	CATTTGCTAGTTTATTCTTGACATGAAGCGTGCCTAATGGTATATTACTTGAGTACTGTT	4.9	2.8
CP4	GATGTTTTAGTTTATTCTTGACACCGTATCGTGCGCGTGATATAATCGGGATCCTTAAGA	5.1	1.3
CP44=CP45	CATCGGGTAGTTTATTCTTGACA-ATTAAGTAGAGCCTGATATAATAGTTCAGTACTGTT	12	34
CP1	CATACCGGAGTTTATTCTTGACAGTTCCACCTCGGGTTGATATAATATCTCAGTACTGTT	34	3.1
CP19=CP22	CATCGCTTAGTTT-TTCTTGACAGGAGGGATCCGGGTTGATATAATA-GTTAGTACTGTT	58	3.3
CP34	CATCGCGAAGTTTATTCTTCACACACCGCAGAACTTGTGGTATAATACAACAGTACTGTT	59	8.4
CP20	CATGGGTGAGTTTATTCTTGACAGTGCGGCCGGGGGCTGATATCATAGCAGAGTACTATT	60	22
CP11	CATAAGTGAGTTTATTCTTGACCCGGACGCCCCCCTTTGATATAATAAGT-AGTACTGTT	69	1.4
CP26	CATTCTACAGTTTATTCTTGACATTGCACTGTCCCCCTGGTATAATAACTATACATGCAT	72	10
CP3	CATCCTGTAGTTTATTCTTGACACAAGTCGTTAGCTGTGGTATAATAGGAGAGTACTGTT	74	0.9
CP14	CATGACGGAGTTTATTCTTGACACAGGTATGGACTTATGATATAATAAAACAGTACTGTT	81	0.3
CP13	CATGCTTTACTTTATTCTTGACAAAACCACCAGCTTTTGGTATAATACGTGAGAACTGTT	84	1.0
CP40	CATAGAACAGTTTATTCTTGACATTGAATAAGAAGGCTGATATAATAGC-CAGTACTGTT	104	19
CP8	CATTCTTTAGTTTATTCTTGACAAACGTATTGAGGACTGATATAATAGGTGAGTACTGTT	146	1.2
CP28	CATGGGGCCGTTTATTCTTGACAACGGCGAGCAGACCTGGTATAATAATATAGTACTGTT	161	0.9
CP10	CATGGCTTAGTTTATTCTTGACAGGGTAGTATCACTGTGATATAATAGGACAGTACTGTT	181	2.5
CP32	CATACGGGAGTTTATTCTTGACATATTGCCGGTGTGTTGGTATAATAACTTAGTACTGTT	214	60
CP30	CATGACAGAGTTTATTCTTGACAGTATTGGGTTACTTTGGTATAATAGTTGAGTACTGTT	228	42
CP9a	CATAGTCTAGTTTATTCTTGACACGCGGTCCATTGGCTGGTATAATAATTTAGTACTGTT	243	56
CP38	CATAGAGAAGTTTATTCTTGACAGCTAACTTGGCCTTTGATATAATACATGAGTACTGTT	244	92
CP46	CATGATGTAGTTTATTCTTGACACTGAGAGGGCCTCTTGATATAATAGTTGAGTACTGTT	256	33
CP23	CATGTAGGAGTTTATTCTTGACAGATTAGTTAGGGGGTGGTATAATATCTCAGTACTGTT	257	2.6
CP39	CATTGCGAAGTTTATTCTTGACAGTACGTTTTTACCATGATATAATAGTATAGTACTGTT	266	0.5
CP33=CP35	CATGTTGGAGTTTATTCTTGACATACAATTACTGCAGTGATATAATAGGTGAGTACTGTT	298	7.0
CP15	CATTACGTAGTTTATTCTTGACAGAATTACGATTCGCTGGTATAATATCAGTACTGTT	322	1.5
CP29	CATCGGTAAG-TTATTCTTGACATCTCAGGGGGGACGTGGTATAATAACTGAGTACTGTT	399	1.4
CP12b	CATATACAAGTTTATTCTTGACACTAGTCGGCCAAAATGATATAATACCTGAGTACTGTT	419	101
CP41	CATCCGCAAGTTTATTCTTGACAGCTGAATGTAGACGTGGTATAATAGTTAAGTACTGTT	624	18
CP16	CATTGTGTAGTTTATTCTTGACAGCTATGAGTCAATTTGGTATAATA--ACAGTACTCAG	627	0.3
CP42	CATTCGTAAGTTTATTCTTGACACCTGAGATGAGGCGTGATATAATAAATAAGTACTGTT	680	7.2
CP7	TATGCGGTAGTTTATTCTTGACATGACGAGACAGGTGTGGTATAATGGGTCTAGATTAGG	1021	134
CP6	CATGTGGGAGTTTATTCTTGACACAGATATTTCCGGATGATATAATAACTGAGTACTGTT	1876	280
CP25	C-TTTGGCAGTTTATTCTTGACATGTAGTGAGGGGGCTGGTATAATCACATAGTACTGTT	2050	528

FIG. 4. Sequence of the area from positions −52 to +8 (relative to the putative transcription initiation site) of the synthetic promoter clones pCP1 through pCP46. The clones are ordered according to strength. Matches to the oligonucleotide consensus sequence (given at the top) are in boldface. Errors in the −35 or −10 consensus sequence and deletions in the spacer between these sequences are underlined. Two clones, CP9 and CP12, had two promoter fragments inserted in tandem, a (upstream fragment) and b (downstream fragment). In these cases, only one of the two tandem promoters was perfect; data for these promoters are shown. β-gal, β-galactosidase.

Regulation of promoter activities. The synthetic CP promoters were designed to be constitutive. To test this experimentally, the expression in exponential growth phase and stationary growth phase was measured for a selection of the promoter clones. We found that the specific activity of β-galactosidase was two- to fourfold higher in the stationary-phase cultures than in the exponential-phase cultures (data not shown). However, the copy number of the vector used in these studies has been shown to increase approximately threefold in the stationary phase (11), which demonstrates that the CP promoters are indeed quite close to being constitutive under these conditions.

Activities of the synthetic promoters in *E. coli*. Another interesting point is whether the promoters are functional in other organisms, and if so, whether the relative strength of the promoters would be dependent on the organism. As described above, the promoter cloning vector, pAK80, that we used here for construction of the synthetic promoters also replicates in *E. coli*; indeed, the promoter clones were first isolated in *E. coli*. We could therefore measure the activities of the synthetic promoters also in *E. coli* (Fig. 5). The promoter strength was also highly variable for the individual promoters in this organism, and we found that the promoters covered activities from 0.2 to 500 Miller units. In this case also, the activity increased in small steps.

The absolute values of β-galactosidase units measured in *E. coli* were lower on average compared to *L. lactis*; this was probably a consequence of a low efficiency of translation of the *lacL* and *lacM* genes in *E. coli*, since these genes and their ribosome binding sites originate from the gram-positive bacterium *Leuconostoc mesenteroides*. When some of the strongest promoters were cloned into a promoter cloning vector designed for *E. coli*, the promoters turned out to be quite strong (data not shown).

Figure 6 shows a plot of activity of the CP promoters in *L. lactis* and *E. coli*. The strengths of the individual CP promoters in the two organisms correlate somewhat but not very well: some promoters which were quite strong in *L. lactis* were relatively weak in *E. coli*, and vice versa. Moreover, the pattern that we observed in *L. lactis*, i.e., that the relatively strong promoters were the perfect ones, did not hold true for *E. coli*: here the promoters which had either an error in the consensus sequence or a shorter spacer were relatively strong.

DISCUSSION

We have constructed a library of synthetic promoters that differ in strength over 3 to 4 logs of activity, and this range of activity is covered by small steps of activity increase. Moreover, some of the promoters that resulted from this random approach turned out to be quite strong.

The fact that the library of promoters covered such a wide range of activities was somewhat surprising to us; the underlying idea behind the construction of the CP promoters was that the context of the consensus sequences (the spacers) would play a role in modulating the strength of a promoter, rather than changing the activity over several logs of activity. Indeed, much of that variation (below 5 Miller units) was probably a consequence of the accidental introduction of mutations in the consensus sequences and in the length of the spacer regions. In contrast, the strong promoters in *L. lactis* (those having activities higher than 100 Miller units) were all perfect with respect to the consensus sequence and spacer length. But even when we confine our analysis to these promoter clones, we find 400-fold variation in promoter activity, still in small steps of activity increase, which demonstrates that the context in which the consensus sequences are embedded (i.e., the spacers) clearly is important for promoter strength.

The ranking of the promoters depended on the organism in

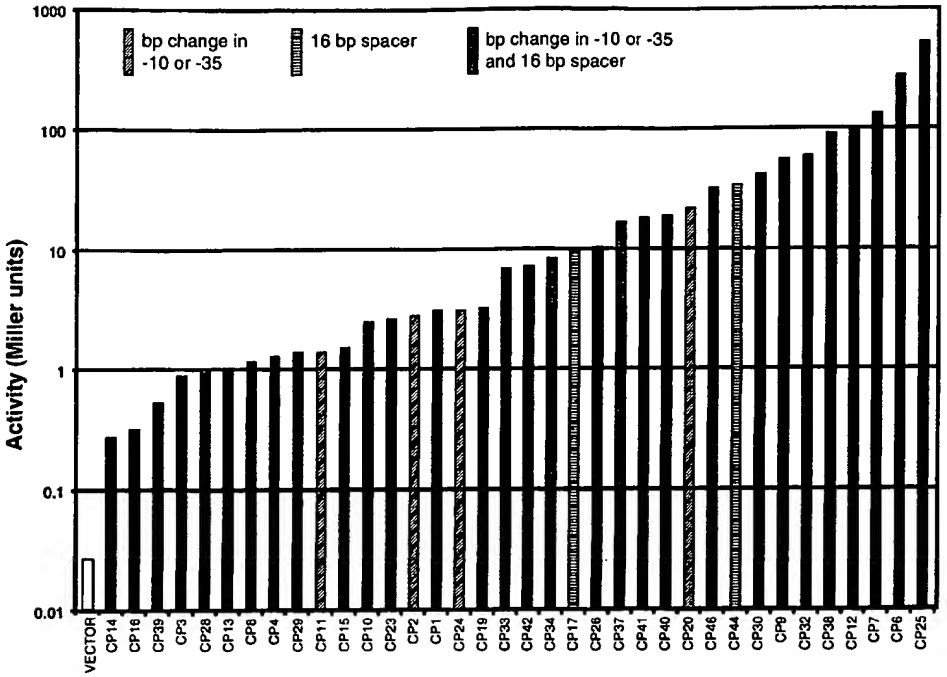

FIG. 5. β-Galactosidase activities of the CP promoters in *E. coli*. The promoter activities were assayed from the expression of a reporter gene (*lacLM*) encoding β-galactosidase transcribed from the different synthetic promoter clones on the promoter cloning vector pAK80. The patterns of the data points indicate which promoter clones contained errors in either the −35 or the −10 consensus sequence or in the length of the spacer between these sequences. See text for further details.

which they were measured, possibly because the σ factor-RNA polymerase complexes that recognize these promoters have different structures in the two organisms due to differences in amino acid sequences. The fact that *E. coli* accepted some of the less perfect CP promoters as relatively strong promoters could indicate that *E. coli* is more promiscuous with respect to promoter structure than *L. lactis*. This makes some sense considering the composition of the *L. lactis* genome: the AT content is 65%, which is much closer to the base composition of the −35 and −10 consensus sequences. These sequences are therefore more likely to accidentally occur in *L. lactis*, and a stricter requirement for promoter sequences might therefore be expected for this organism.

The process of transcription initiation consists of several events (reviewed in reference 17). First, recognition and binding of the σ factor-RNA polymerase complex to the promoter region takes place (closed complex formation). Subsequently, there is local melting of the DNA double helix (open complex formation), possibly assisted by local negative DNA supercoiling. Finally, the binding between the σ factor-RNA polymerase complex and the promoter area must dissociate and clear the promoter area, so that another initiation complex may form. From this model, it is clear that efficient binding between the σ factor-RNA polymerase complex and the promoter area does not guarantee a strong promoter; promoter strength must be a compromise between binding, melting, and clearance, and probably other factors as well.

What then controls the strength of the individual synthetic promoters presented here? It does not appear that any additional conserved sequence motifs have been generated among the strongest promoters. Rather, it seems that the overall three-dimensional structure which arises from a particular nucleotide sequence could be important.

The method presented here for tuning gene expression in the living cell has both advantages and disadvantages compared to the methods that would use an inducible expression system such as the *lac* promoter. A disadvantage is that instead of only one genetic construct, perhaps three to four constructs have to be made. On the other hand, the constructs are made

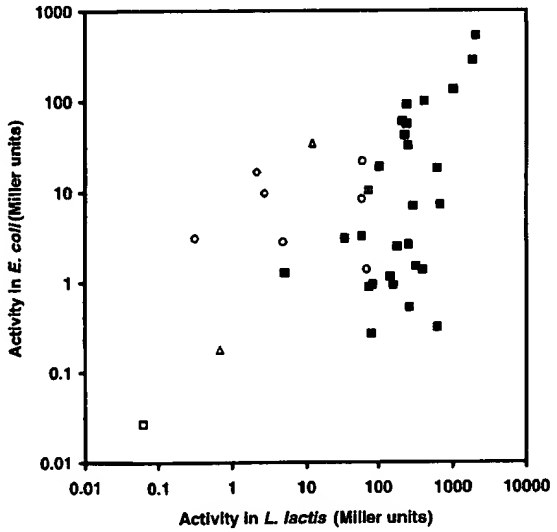

FIG. 6. Correlation between promoter activities in *L. lactis* and *E. coli*. The promoter activities measured in *E. coli* (from Fig. 5) were plotted as a function of the promoter activities measured in *L. lactis* (from Fig. 3). The symbols indicate errors in either the −35 or −10 sequence (solid circles), a 16-bp spacer (triangles), or promoters with both of these errors (diamonds). The open square represents the vector clone.

in parallel, so that the amount of work should not be proportional to the number of constructs. The inducible systems have the advantage that gene expression can be turned on at the proper time during a fermentation, which is sometimes essential (for instance, when the product is toxic to the host cell). The work presented here was aimed at generating a library of constitutive promoters, for achieving a constant level of gene expression throughout the growth of a culture. We are currently working on synthetic inducible promoters in which a regulatory motif has been added. This should allow us to generate libraries of promoters, which differ in basal expression level and can be induced to various extents, by changing a fermentation parameter (i.e., temperature, pH, or salt concentration) or by adding a specific inducer.

The system presented here also has advantages. One is that it is easier to attain a steady expression level of the enzyme in question, which is often quite difficult with inducible systems such as the *lac* system (8). With the method presented here, once the optimal expression level of the enzyme has been determined, the engineered strain is ready to use directly in the fermentation process.

An important feature of the system described here, in a longer perspective, is the possibility to simultaneously modulate, to different extents, the expression of several individual genes or operons located at various positions of the genome in the same strain. Metabolic control analysis (5, 10) showed that in theory, flux and concentration control can be shared among several enzymes in a pathway, and experimental determinations of flux control have often showed that control seems to be distributed over many enzymes in the living cell (9, 15, 18, 19, 22, 23): in most cases, there may not be such a thing as a rate-limiting step, and even if one finds a step that has a measurable control, the control will often disappear relatively quickly as the enzyme is being overexpressed. Since the sum of flux control must equal unity, this then means that flux control has been shifted to other steps in the pathway. In summary, in order to increase a given flux in a living cell, it may thus be necessary to (i) optimize the individual expression of several genes and (ii) after one round of optimization in which one enzyme was clamped at the optimal level, continue the optimization of other enzymes in the pathway. With the systems available until now, one would then quickly run out of expression systems to use, but with our method, one can in principle continue the optimization numerous times.

In this report, the method for generating synthetic promoters of different strengths was illustrated for use in the gram-positive bacterium *L. lactis*. However, there is no obvious reason why the approach should be limited to this organism, and the fact that the same promoter library was also functional in the gram-negative bacterium *E. coli* suggests that the approach may be universally applicable to prokaryotic organisms. An exciting question is then, can the approach be extended to work for modulating gene expression in eukaryotic cells? Such experiments are under way, and the results are quite encouraging.

ACKNOWLEDGMENTS

We are deeply indebted to Regina Schürmann for excellent technical assistance.

This work was funded by the Danish Centre for Advanced Food Studies.

REFERENCES

1. **Boe, L.** Personal communication.
1a. **Casabadan, M. J., and S. N. Cohen.** 1980. Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli*. J. Mol. Biol. **138**:179–207.
2. **de Ruyter, P. G., O. P. Kuipers, and W. M. de Vos.** 1996. Controlled gene expression systems for *Lactococcus lactis* with the food-grade inducer nisin. Appl. Environ. Microbiol. **62**:3662–3667.
3. **de Vos, W. M., and G. Simons.** 1994. Gene cloning and expression systems in lactococci, p. 52–105. *In* M. J. Gasson and W. M. de Vos (ed.), Genetics and biotechnology of lactic acid bacteria. Blackie Academic & Professional, Glasgow, United Kingdom.
4. **Gasson, M. J.** 1983. Plasmid complements of *Streptococcus lactis* NCDO 712 and other lactic streptococci after protoplast-induced curing. J. Bacteriol. **154**:1–9.
5. **Heinrich, R., and T. A. Rapoport.** 1974. A linear steady-state treatment of enzymatic chains: general properties, control and effector-strength. Eur. J. Biochem. **42**:89–95.
6. **Holo, H., and I. F. Nes.** 1989. High-frequency transformation, by electroporation, of *Lactococcus lactis* subsp. *cremoris* grown with glycine in osmotically stabilized media. Appl. Environ. Microbiol. **55**:3119–3123.
7. **Israelsen, H., S. M. Madsen, A. Vrang, E. B. Hansen, and E. Johansen.** 1995. Cloning and partial characterization of regulated promoters from *Lactococcus lactis* Tn*917-lacZ* integrants with the new promoter probe vector, pAK80. Appl. Environ. Microbiol. **61**:2540–2547.
7a. **Jensen, P. R.** 1997. International patent application PCT/DK97/00342.
8. **Jensen, P. R., H. V. Westerhoff, and O. Michelsen.** 1993. The use of *lac*-type promoters in control analysis. Eur. J. Biochem. **211**:181–191.
9. **Jensen, P. R., H. V. Westerhoff, and O. Michelsen.** 1993. Excess capacity of H^+-ATPase and inverse respiratory control in *Escherichia coli*. EMBO J. **12**:1277–1282.
10. **Kacser, H., and J. A. Burns.** 1973. The control of flux. Symp. Soc. Exp. Biol. **27**:65–104.
11. **Madsen, P. L.** 1996. Transcription of the lactococcal temperate phage TP901-1. Ph.D. thesis. Department of Biological Chemistry, University of Copenhagen, Copenhagen, Denmark.
12. **Makrides, S. C.** 1996. Strategies for achieving high-level expression of genes in *Escherichia coli*. Microbiol. Rev. **60**:512–538.
13. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor, Cold Spring Harbor Laboratory, N.Y.
14. **Miller, J. H.** 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
15. **Niederberger, P., R. Prasad, G. Miozzari, and H. Kacser.** 1992. A strategy for increasing an in vivo flux by genetic manipulations. Biochem. J. **287**:473–479.
16. **Nilsson, D., and E. Johansen.** 1994. A conserved sequence in tRNA and rRNA promoters of Lactococcus lactis. Biochim. Biophys. Acta **1219**:141–144.
17. **Pérez-Martín, J., F. Rojo, and V. de Lorenzo.** 1994. Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. Microbiol. Rev. **58**:268–290.
18. **Ruijter, G. J. G., P. W. Postma, and K. van Dam.** 1991. Control of glucose metabolism by enzyme II^{Glc} of the phosphoenolpyruvate-dependent phosphotransferase system in *Escherichia coli*. J. Bacteriol. **173**:6184–6191.
19. **Schaaff, I., J. Heinisch, and F. K. Zimmermann.** 1989. Overproduction of glycolytic enzymes in yeast. Yeast **5**:285–290.
20. **Schickor, P., W. Metzger, W. Werel, H. Lederer, and H. Heumann.** 1990. Topography of intermediates in transcription initiation of *E. coli*. EMBO J. **9**:2215–2220.
21. **Schneider, K., and C. F. Beck.** 1986. Promoter-probe vectors for the analysis of divergently arranged promoters. Gene **42**:37–48.
22. **Snoep, J. L., L. P. Yomano, H. V. Westerhoff, and L. O. Ingram.** 1995. Protein burden in *Zymomonas mobilis*: negative flux and growth control due to overproduction of glycolytic enzymes. Microbiology **141**:2329–2337.
23. **Walsh, K., and D. E. Koshland, Jr.** 1985. Characterization of rate-controlling steps in vivo by use of an adjustable expression vector. Proc. Natl. Acad. Sci. USA **82**:3577–3581.

JOURNAL OF BACTERIOLOGY, Oct. 1995, p. 5740–5747
0021-9193/95/$04.00+0

Vol. 177, No. 20

Nucleotide Sequence, Transcriptional Analysis, and Glucose Regulation of the Phenoxazinone Synthase Gene (*phsA*) from *Streptomyces antibioticus*

CHUIN-JU HSIEH AND GEORGE H. JONES*

Department of Biology, Emory University, Atlanta, Georgia 30322

Received 27 June 1995/Accepted 8 August 1995

The nucleotide sequence of a 2.3-kb *Sph*I fragment containing the structural gene (*phsA*) for phenoxazinone synthase (PHS) of *Streptomyces antibioticus* was determined. The sequence was found to contain an open reading frame (ORF) with a G+C content of 71.5% oriented in the direction of transcription that was confirmed by primer extension. The ORF encodes a protein with an M_r of 70,223 consisting of 642 amino acids and is preceded by a potential ribosome-binding site. The codon usage pattern is in agreement with the general pattern for streptomycete genes, with a 92.5 mol% G+C content in the third position. The N-terminal sequence of the mature PHS subunit corresponds exactly to that predicted from the nucleotide sequence. Neither ATG nor GTG initiator codons were identified for the protein. However, a TTG codon was located near the amino terminus of the mature protein and is a good candidate for the initiator codon. The transcriptional start point of *phsA* was located 36 bp upstream of the start codon by primer extension. The −10 region of the putative promoter showed some similarity to the consensus sequence for the major class of prokaryotic promoters, but the −35 region was less similar. Comparison of the primary amino acid sequence of PHS of *S. antibioticus* with other amino acid sequences indicated that PHS is a blue copper protein with copper binding domains in the N-terminal and C-terminal regions of the polypeptide chain. A *Bsr*BI fragment containing the promoter region of *phsA* and a portion of the ORF was shown to promote *xylE* expression when cloned in the streptomycete promoter probe vector pIJ2843. This *phsA* promoter-dependent *xylE* expression could be repressed by glucose in *S. antibioticus* when the organism was grown on glucose or galactose plus glucose. Thus, the cloned promoter region appears to contain the sequences responsible for catabolite repression of PHS production.

Actinomycin is one of the antibiotics produced by the gram-positive actinomycete *Streptomyces antibioticus* (52). A putative pathway for actinomycin biosynthesis was proposed several years ago (50), and biochemical, physiological, and genetic studies have confirmed the essential details of that pathway. Five enzymes from *S. antibioticus*, *Streptomyces chrysomallus*, and *Streptomyces parvulus* have been isolated and characterized to demonstrate their involvement in the actinomycin biosynthetic pathway (6, 11, 22, 23, 28–31). One of these enzymes, phenoxazinone synthase (PHS), catalyzes the oxidative condensation of two molecules of 4-methyl 3-hydroxyanthraniloyl pentapeptide to form actinomycinic acid, which is the penultimate intermediate in the putative biosynthetic pathway (Fig. 1). The enzyme was first identified by Katz and Weissbach (28) and subsequently purified by Choy and Jones (6). To date, *phsA*, the gene coding for PHS from *S. antibioticus*, is the only gene involved in actinomycin biosynthesis that has been cloned (25).

Although essentially all of the enzymes required for actinomycin production have been identified, little is known about the regulation of these enzymes and of overall actinomycin production. Of all the enzymes identified, PHS is perhaps the best characterized. It was shown some years ago by Marshall and coworkers that actinomycin production is repressed in *S. antibioticus* cultures grown on glucose or galactose plus glucose as compared with cultures grown on production medium with galactose alone as the carbon source (39).

Catabolite control has been implicated in the expression of both PHS and actinomycin synthetase I (ACMSI; the enzyme which activates the precursors of the actinomycin chromophore in *S. antibioticus* [23, 30, 31]). PHS production was demonstrated to be subject to catabolite control shortly after the identification of the enzyme (13, 28). It is possible that *phsA* and *acmsI* are located in the same genomic region in *S. antibioticus* since it was well known that the genes for antibiotic production are clustered in the streptomycete genome (for examples, see reference 38). Therefore, the detailed molecular analyses of the mechanisms controlling the expression of *phsA* are essential to our understanding of the regulation of actinomycin biosynthesis and the synthesis of other antibiotics. We report here the nucleotide sequence and transcriptional analysis of *phsA* and identify the promoter region of the gene. We also demonstrate that a cloned fragment containing the putative promoter is active in a streptomycete promoter probe vector and that the activity of the promoter is repressed when *S. antibioticus* transformants containing the relevant constructs are grown on glucose or galactose plus glucose as compared with cultures grown on galactose as the sole carbon source.

MATERIALS AND METHODS

Organisms and growth conditions. The *Streptomyces* strains used were *S. antibioticus* IMRU 3720 and *Streptomyces lividans* 66 derivative TK24 (18). *S. antibioticus* was grown on liquid NZ-amine and galactose-glutamic acid media as described previously (13). *S. lividans* was generally grown on yeast extract-malt extract plus 34% sucrose (YEME) or on tryptone soy broth. For protoplast preparation, TK24 was grown on YEME with $MgCl_2$ and glycine at the final concentrations of 5 mM and 0.5%, respectively (49). Protoplasts were allowed to regenerate on R2YE medium (49) for 12 to 24 h and then overlaid with 2 to 3 ml of soft nutrient agar supplemented with thiostrepton at a final concentration of 500 μg/ml. Tyrosine at 0.075% (wt/vol) was added to the soft nutrient agar for overlaying when pIJ702 derivatives were used.

Escherichia coli DH5α [F⁻ φ80 *dlacZΔM15* (*lacZYA-argF*) *U169 endA1 recA1 hsdR17* (r_K^+ m_K^+) *deoR thi-1 supE441 gyrA96 relA1*] and XL-1 Blue 2 [*recA1*

* Corresponding author. Phone: (404) 727-4208. Fax: (404) 727-2880. Electronic mail address: GJONES@BIOLOGY.EMORY.EDU.

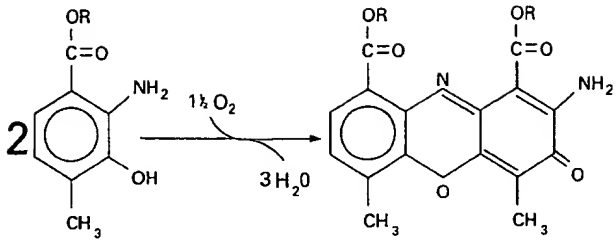

FIG. 1. The PHS reaction. The penultimate step in the actinomycin biosynthetic pathway in *S. antibioticus*, the oxidative condensation of two molecules of 4-methyl 3-hydroxyanthraniloyl pentapeptide to yield actinomycinic acid, is catalyzed by PHS.

endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac (F′ *proAB lacI*^{q}*Z M15* Tn*10* (Tet^r)] were generally cultured in L broth or on L agar (35). *E. coli*-competent cells were prepared by the $CaCl_2$ method and transformed as described by Sambrook et al. (46). After transformation with pUC19 and pBluescript SK$^+$ derivatives, transformants were selected on L agar plates containing 100 μg of ampicillin per ml, 40 mg of 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal) per ml, and 0.2 mM isopropyl-β-D-thiogalactopyranoside (IPTG). For single-stranded DNA preparation, strains containing pBluescript SK$^+$ derivatives were grown in 2XYT medium in the presence of 100 μg of ampicillin per ml and helper phage VCSM13 for 2 h followed by the addition of 75 μg of kanamycin per ml and growth overnight. Growth temperatures for *Streptomyces* spp. and *E. coli* were 30 and 37°C, respectively.

DNA manipulations. Plasmid and chromosomal DNAs were prepared as described previously (4, 17, 20) and analyzed by restriction digestion and agarose gel electrophoresis. In some experiments, restriction fragments were recovered from low-melting-point agarose as described by Favre (10). Protoplast preparation, transformation, and regeneration were as described previously (17, 20, 25). A list of plasmids used or generated in the present study is provided in Table 1. pJSE923 is a derivative of pIJ2501 (25) with an *Xba*I linker inserted at the *Pvu*I site of *phsA*. pJSE929 contains the blunt-ended *Bsr*BI subfragment of the *phsA* promoter region cloned into the *Hin*cII site of pUC19. pJSE935 contains the *Hin*dIII-*Bam*HI subfragment of the *phsA* promoter region of pJSE929 cloned into *Hin*dIII-*Bam*HI-digested pIJ2843 (7).

Enzyme assays. *Streptomyces* cultures were grown in 250-ml flasks containing 50 ml of glutamic acid-salts medium, 50 μg of thiostrepton per ml as necessary, and 5 mM $CuSO_4$ at 28°C with shaking at 200 rpm. Cultures contained either 1% galactose, 1% glucose, or 0.5% galactose plus 0.5% glucose as carbon sources. The cultures were harvested 12 h after inoculation. Mycelium was washed in 100 mM potassium phosphate (pH 7.5), suspended in a final volume of 2 ml of sample buffer (19), and disrupted by sonication.

Catechol dioxygenase assays were performed and activities were determined spectrophotometrically as described previously (19, 54). Catechol dioxygenase specific activity was calculated as the rate of change in A_{375} per min per milligram of protein and converted to milliunits per milligram (45). Protein concentrations were determined with the bicinchoninic acid protein assay reagent kit from Pierce. The PHS assay was performed as described previously (6) with 3-hydroxyanthranilic acid as the substrate.

Nucleotide sequence analysis. Sequential deletion clones from both ends of the *phsA Sph*I fragment were obtained by exonuclease III-mung bean nuclease digestion with the exonuclease III-mung bean deletion kit from Stratagene Cloning Systems. The *phsA Sph*I fragment was subcloned into pBluescript SK$^+$ (Stratagene) modified to contain an *Sph*I site in the polylinker, and the resulting recombinant plasmids (pJSE900 and pJSE910) were used to create deletion clones suitable for sequencing. The nucleotide sequences of both DNA strands of the cloned *phsA* fragment were obtained by the dideoxy chain termination method (47). Single-stranded DNA was obtained with VCSM13 as a helper phage, and the DNA was prepared as described previously (26). The sequencing reactions were performed basically as described for the 7-deaza-GTP Sequenase kit from United States Biochemicals except that the extension and termination reactions were done at 50 and 70°C, respectively. The reactions were post-terminated at 70°C for 2.5 min by adding 2.5 U of *Taq* version 2.0 DNA polymerase and 1 μl of termination mixture, both from United States Biochemicals. Difficult compression areas and pause sites were resolved by using dITP instead of deaza-GTP. The DNA sequences were analyzed with the DNAsis program from Hitachi and the GCG program from the University of Wisconsin.

The GenBank accession number for the *S. antibioticus* IMRU3720 PHS gene (*phsA*) is U04283.

Primer extension. In the primer extension experiments, a 24-base oligonucleotide primer, 5′-GATCTCGGTCTCCCGCGTCACCTC-3′, that is located 528 bp downstream of the 5′-*Sph*I site and is complementary to the *phsA* mRNA was used to reveal the transcriptional start point. End labeling of the 5′-terminus of the oligonucleotide primer with the polynucleotide kinase reaction and the primer extension reaction were done as described by Moran (42). RNA preparation was as described previously (17) with the following modifications. Mycelium was collected on a Whatman no. 4 filter disc by use of a vacuum line to accelerate the filtration process. The mycelium was quickly scraped off the filter into a universal bottle and resuspended in 5 ml of modified Kirby mixture at 4°C (modified Kirby mixture consists of 1% [wt/vol] sodium triisopropylnaphthalene sulfonate [Eastman Chemicals], 6% [wt/vol] sodium-4-amino salycilic acid [sodium salt; BDH], and 6% [vol/vol] Tris-EDTA-buffered phenol mixture, and all solutions were made up in 50 mM Tris-HCl [pH 8.3]). The contents were vortexed with 10 g of 4.5- to 5.5-mm-diameter glass balls as vigorously as possible for at least 2 min. Three milliliters of phenol-chloroform mixture was added, and the mixture was vortexed as described above. The homogenate was then transferred to a polypropylene tube (Falcon 2006) and centrifuged (10 min at 12,000 × *g* and 4°C) to separate the phases. The aqueous layer was transferred to a fresh tube, and an additional 5 ml of phenol-chloroform mixture was added. The solutions were vortexed thoroughly for 2 min and centrifuged again as described previously to separate the phases, and this procedure was repeated until very little interphase material remained visible. One-tenth volume of 4 M sodium acetate (pH 6.0), followed by an equal volume of isopropanol, was added to the aqueous phase. The solutions were mixed and left at −20°C for 1 h. The nucleic acids were collected by centrifugation at 12,000 × *g* for 10 min, and the supernatant was discarded. The pellet was rinsed with absolute ethanol and vacuum dried. The pellet was resuspended in 180 μl of distilled water (treated with diethyl pyrocarbonate) and 20 μl of 10× DNase buffer (0.5 M Tris-HCl [pH 7.8], 0.05 M $MgCl_2$) and transferred to an Eppendorf tube. DNase (RNase-free; Sigma Chemical Co.) was added to a final concentration of 30 μg/ml. The solutions were incubated at room temperature for 30 min. An equal volume of phenol-chloroform mixture was then added, and the samples were mixed by vortexing. The phases were separated by centrifugation in a microcentrifuge, and the aqueous phase was transferred to a fresh tube. The aqueous phases were then extracted by adding an equal volume of chloroform. Total RNA was precipitated with 1/10 volume of 3 M sodium acetate (pH 6) and an equal volume of isopropanol for 2 h at −20°C, and the precipitate was collected by centrifugation. The RNA pellet was rinsed with 70% and then 100% ethanol, vacuum dried, resuspended in 100 μl of distilled water, and stored at −70°C. The quantity of RNA was assessed by spectrophotometry, and the quality was assessed by agarose gel electrophoresis.

TABLE 1. Plasmids used or referred to in the present study

Plasmid	Description	Source or reference
pUC19		53
pBluescript SK$^+$	Phagemid cloning vector (Stratagene); the vector was modified to contain an *Sph*I site in the polylinker	
pIJ702		27
pIJ2501	The 2.3-kb *phsA Sph*I structural gene from *S. antibioticus* cloned into the *Sph*I site of pIJ702	25
pIJ2843	*Streptomyces* low-copy-number promoter-probe vector	7
pJSE900 and pJSE910	The 2.3-kb *phsA Sph*I cloned in the *Sph*I site of pBluescript SK$^+$ in two orientations	Present study
pJSE923	pIJ2501 with an *Xba*I linker at the *Pvu*I site of *phsA*	Present study
pJSE929	The blunt-ended, ca. 235-bp *Bsr*BI subfragment of the *phsA* promoter region, extending from position −106 to +135 relative to the transcriptional start site, cloned into the *Hin*cII site of pUC19	Present study
pJSE935	The ca. 265-bp *Hin*dIII-*Bam*HI subfragment of the *phsA* promoter region of pJSE929 cloned into *Hin*dIII-*Bam*HI-digested pIJ2843	Present study

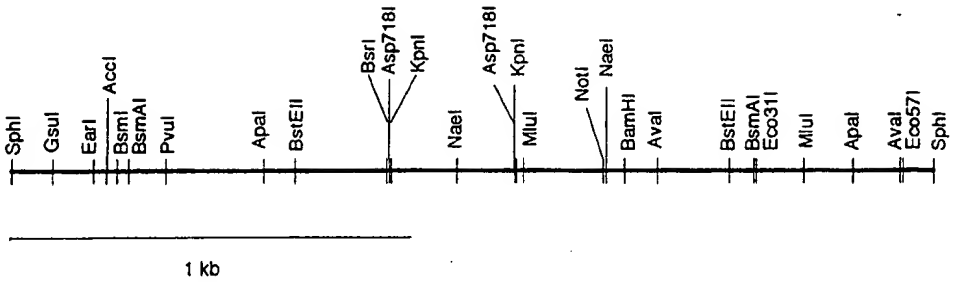

FIG. 2. Restriction map of *phsA* constructed from the *phsA* sequence.

Determination of the amino-terminal sequence of the PHS subunit. The amino-terminal sequences of the cloned and native PHS proteins were determined at the Emory University Microchemical Facility and found to be identical. The sequence of the first 15 amino acids of the protein is Thr-Asp-Met-Ile-Glu-Gln-Ser-Asp-Asp-Arg-Ile-Asp-Pro-Ile-Asp.

Enzymes and reagents. Restriction endonucleases were purchased from Boehringer-Mannheim Corporation, Gibco BRL, and Promega Corporation. Calf intestinal alkaline phosphatase, T4 DNA ligase, and avian reverse transcriptase were obtained from United States Biochemicals. Exonuclease III and mung bean nuclease were obtained from Stratagene. Sigma Chemical Co. supplied RNase, which was prepared as described previously (46). The 7-deaza-dGTP Sequenase version 2.0 and *Taq* version 2.0 DNA polymerase kits were purchased from United States Biochemicals. [τ-^{32}P]dATP, [α-^{32}P]dCTP, and α-^{35}S-dATP were purchased from Dupont New England Nuclear Products and Amersham. RNasin was obtained from Promega. All of the chemicals were of reagent grade or the highest purity commercially available.

RESULTS

Nucleotide sequence analysis. A detailed restriction map of the *phsA SphI* fragment constructed on the basis of the nucleotide sequence is shown in Fig. 2, and the nucleotide sequence of the fragment is shown in Fig. 3. Analysis of the DNA sequence with the FRAME codon preference program (3) revealed a 1,932-bp open reading frame with 71.5% G+C content, matching the codon usage of *Streptomyces* spp. (Fig. 4). The open reading frame presumably starts with a TTG codon at nucleotide 348 and encodes a deduced polypeptide of 642 amino acids with a predicted M_r of 70,223. Furthermore, the predicted initiator amino acid is only one position upstream of the N-terminal amino acid obtained by protein sequence analysis of purified PHS (the first 15 amino acids shown in Fig. 3; see Materials and Methods). Additional information on the putative translational start was obtained by inserting an *Xba*I linker downstream of this region (Table 1, pJSE923). The inserted linker created stop codons in all three reading frames. When the resulting recombinant plasmid, containing the *Xba*I linker in the *Pvu*I site of *phsA*, was used to transform *S. lividans*, PHS expression from *phsA* was completely abolished (data not shown). These results rule out the possibility that the cloned fragment activates a normally silent *phsA* gene in *S. antibioticus*, as has been observed for *S. lividans* (25, 37). Upstream of the putative TTG start codon is the sequence GGGGG (Fig. 3, boxed), which may act as a ribosome binding site (48). A short stem-loop structure is located 4 bp downstream of the *phsA* stop codon (Fig. 3, inverted arrows), but its ability to function in transcription termination is problematic because of its length.

Primer extension analysis and identification of the putative *phsA* promoter. A 24-mer oligonucleotide primer, corresponding to sequences 530 bp downstream of the 5′ *Sph*I site and 180 bp downstream of the translational start codon (Fig. 3), was used in primer extension studies to locate the 5′ end of the *phsA* transcript (Fig. 5). RNA templates were prepared from *S. antibioticus* and *S. lividans* as indicated in the legend to Fig. 5. The transcriptional start point (*tsp*) of the *phs* message revealed by this analysis is located at the A residue which is 313 bp downstream of the 5′ *Sph*I site and 36 bp 5′ to the translation initiation codon. The transcription start point of the cloned *phsA* gene in *S. lividans* TK24 is the same as that of the chromosomal gene in *S. antibioticus* (Fig. 5). In addition, there is no difference in the *tsp* shown in the primer extension experiments using total RNA prepared from glucose- or galactose-grown cultures (data not shown). However, glucose-grown cultures contained less *phs*-specific message than galactose-grown cultures. This observation is consistent with earlier data suggesting that the decreased level of PHS observed in cultures grown on glucose as compared with that in galactose-grown cultures is due in part to an effect at the level of *phs* transcription (20, 21).

On the basis of primer extension studies, putative −10 and −35 promoter regions were located relative to the transcription start point (Fig. 3). There are also other interesting features which are located near the promoter region, including several sets of direct repeat sequences, two sets of inverted repeats, and two TNTNAN sequences (Fig. 3). These sequences are noteworthy because they may be involved in the catabolite control of the *phsA* gene (41). The function of these sequences will be examined in detail in subsequent studies.

Confirmation of the presence of a functional promoter upstream of the transcription start site was obtained by promoter probe cloning. In these experiments, a *Bsr*BI fragment from *phsA* (see Fig. 2 and 3) was inserted upstream of the *xylE* gene in the promoter probe vector pIJ2843 (7, 36). The resulting recombinant plasmid was used to transform *S. antibioticus* and *S. lividans*, and mycelial extracts were prepared after 19 h of growth of control and transformed cultures in liquid media. The results of catechol dioxygenase assays of those extracts revealed that the untransformed strains contained negligible levels of enzyme activity, as was also the case for strains transformed with pIJ2843. In contrast, *S. antibioticus* and *S. lividans* strains containing pJSE935, with the putative promoter fragment, showed significant levels of *xylE* activity (data not shown). Thus, the *Bsr*BI fragment does possess promoter activity, and the promoter probe results support the identification of the promoter region of *phsA* suggested by the sequencing and primer extension studies. The use of pJSE935 in studies of glucose repression of *phsA* is described below.

Sequence comparisons with PHS sequence. The deduced amino acid sequence of PHS was compared with entries in protein databases provided by GenBank by use of the FASTA program. The sequence with the greatest homology to PHS was that of bilirubin oxidase from *Myrothecium verrucaria* (32). There is 26% identity and 45% similarity between the sequences of PHS and the bilirubin oxidase protein. A lower homology (18% identity, 40% similarity) was found for the

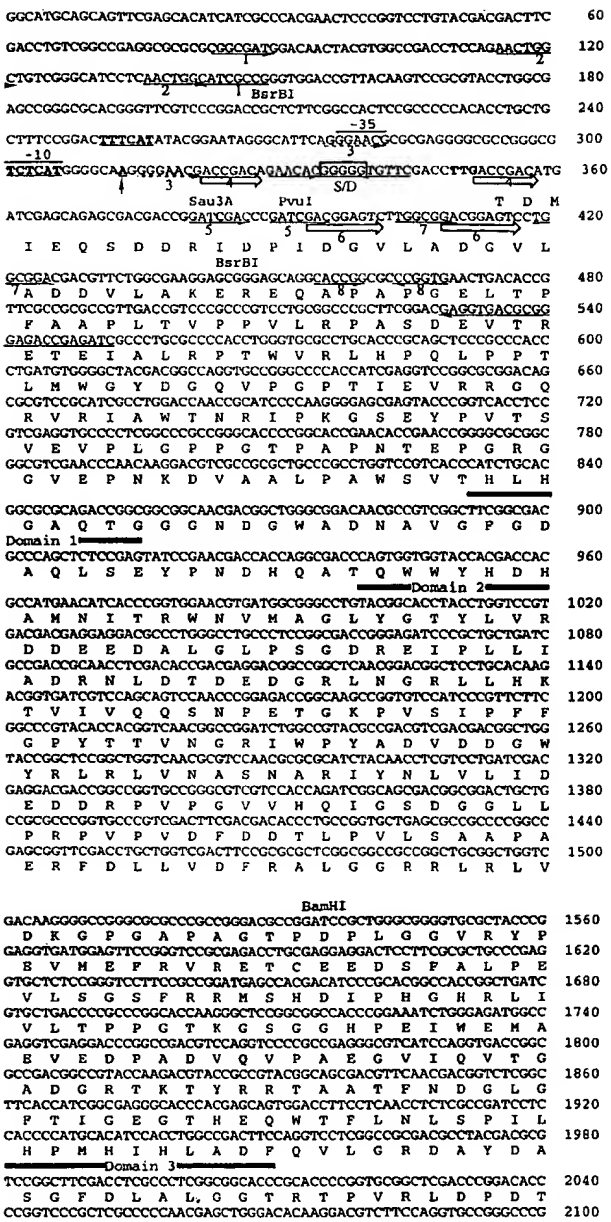

FIG. 3. Nucleotide sequence of the *Sph*I fragment containing the *S. antibioticus phsA* gene. Important restriction enzyme sites are indicated above the sequence. The boxed region denotes the possible ribosomal binding site (Shine-Dalgarno sequence [S/D]). The potential stem-loop is indicated by a pair of inverted arrows located toward the end of the sequence. Potential −10 and −35 regions have lines above the sequence and are discussed in the text. The location of the transcription start point is indicated by an upward arrow and bold type, and the primer for primer extension is shown by the long arrow under the nucleotide sequence around position 530. The direct and inverted repeat sequences are indicated by arrows and numbered in pairs (1 to 8). The TNTNAN elements are shown in bold type and are located around 253 and 303 bp from the 5′ *Sph*I site. Four potential copper binding domains are also indicated in the sequence (solid bars). These sequence data appear in the EMBL, GenBank, and DDBJ nucleotide sequence data libraries under accession number U04283.

sequence of manganese-oxidizing protein from *Leptothrix discophora* (8). Copper binding motifs of all three proteins are aligned in Fig. 6. All three proteins are involved in oxidation reactions, but only PHS and bilirubin oxidase belong to the family of blue copper proteins (2, 12, 32). Sequence comparison of PHS with bilirubin oxidase, manganese-oxidizing protein, and several other blue copper proteins revealed the presence of four regions in the sequence of the former protein corresponding to the potential copper binding domains found in the sequences of the blue copper proteins (Fig. 6). The finding of these copper binding domains confirms PHS as a blue copper protein (2, 12). This result is not at all surprising, since PHS has been shown to require copper for activity (2). The amino acid sequence of PHS contains consensus domains for the copper binding regions of the same types (I, II, and III), which were revealed by X-ray crystallography of ascorbate oxidase from zucchini (40). However, there are just two copper binding domains found in the manganese-oxidizing protein. We speculate that the copper binding domains are components of the catalytic sites of these enzymes.

Expression of the cloned *phsA* promoter is repressed by glucose in *S. antibioticus*. The production of PHS in *S. antibioticus* was demonstrated some years ago to be subject to catabolite control (13, 28). As has been mentioned, later studies suggested that the production of PHS was regulated at the transcriptional level (20, 21). In the present study, the effects of glucose on the expression of the promoter active fragment cloned in pJSE935 was examined in *S. antibioticus*. Transformants containing pJSE935 were grown on 1% galactose, 1% glucose, or a mixture of 0.5% galactose and 0.5% glucose. Catechol dioxygenase assays were performed on extracts of mycelium harvested 12 h after inoculation of the growth media. PHS assays were performed on these same extracts. The results of these experiments, presented in Fig. 7, show that glucose represses the expression of the *phsA* promoter in pJSE935 in the presence or absence of galactose. Thus, the effects of glucose on the *phsA* promoter would seem to fit the classical definition of catabolite repression, which requires that expression of the relevant gene be inhibited when the organism in question is grown on the repressing and (relatively) nonrepressing carbon sources simultaneously. It is significant that the PHS activity in the mycelial extracts exactly paralleled the *xylE* activity; PHS production was inhibited when the organism was grown on glucose alone or on galactose plus glucose (Fig. 7).

One possible mechanism for catabolite repression of *phsA* expression would involve the binding of a repressor protein to operator sequences in the promoter region of the gene. Such mechanisms have been suggested to explain glucose repression in other streptomycetes (for examples, see references 9 and 51). To examine this possibility in *S. antibioticus*, the effects of carbon source on PHS activity were measured in transformants containing pJSE923, in which the *phsA* gene is disrupted by an *Xba*I linker. As controls, PHS activity was measured in untransformed *S. antibioticus* and in transformants containing pIJ702 and pIJ2501. We reasoned that if the *phsA* promoter region possesses a repressor binding site, it might be possible to titrate the repressor by cloning that site at high copy in *S. antibioticus*. However, the result of the experiment was the observation that transformants containing pJSE923 showed the same pattern of PHS expression when grown on galactose, glucose, or glucose plus galactose as did the wild-type strain (data not shown). Thus, although the *Xba*I linker effectively prevented expression of the cloned *phs* gene, the presence of the disrupted gene at high copy did not abolish glucose repression of the endogenous *phsA* gene.

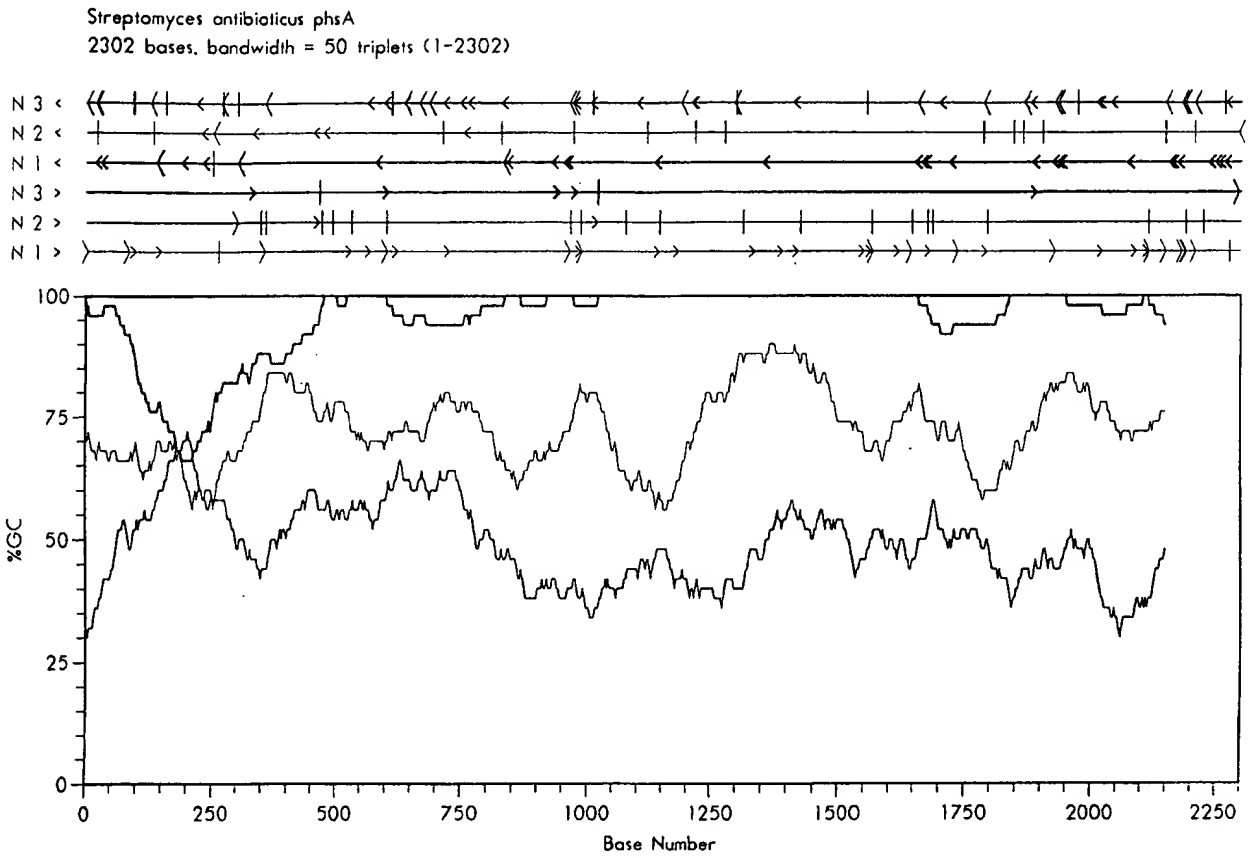

FIG. 4. Analysis of the DNA sequence with the FRAME program (3) revealed a 1,932-bp open reading frame matching the codon usage of *Streptomyces* spp.

DISCUSSION

In the present study, we have characterized the cloned PHS gene from *S. antibioticus*. The M_r of the PHS subunit deduced from the nucleotide sequence data is 70,223. This value differs from the apparent value of 88,000 estimated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) (25). The explanation for the anomalous migration of this protein on SDS-PAGE is not clear. Previous studies have not revealed the presence of carbohydrate or other substances covalently associated with PHS, but other features of the protein presumably cause it to migrate in an unexpected fashion. Two native forms of PHS, large and small (L and S), were reported previously to have M_rs of 540,000 and 180,000 and to be composed of six and two PHS subunits, respectively (6). On the basis of the deduced M_r of the PHS subunit, the corresponding values for L and S would be about 420,000 and 140,000, respectively.

The results of promoter probe cloning and nucleotide sequence analysis of the putative *phsA* promoter support the identity of the −10 and −35 regions and the transcriptional start point of *phsA*. Only a single start point was observed in the experiments illustrated in Fig. 5 and their replicates. It is possible that the *tsp* identified here is artifactual, but it is significant that the use of that start point identifies −10 and −35 regions with significant homology to the P2 promoter of the agarase gene (reference 5 and unpublished results). The −10 region, TCTCAT, of the *phsA* promoter showed more similarity to the −10 consensus sequence, TATAAT, of *E. coli*

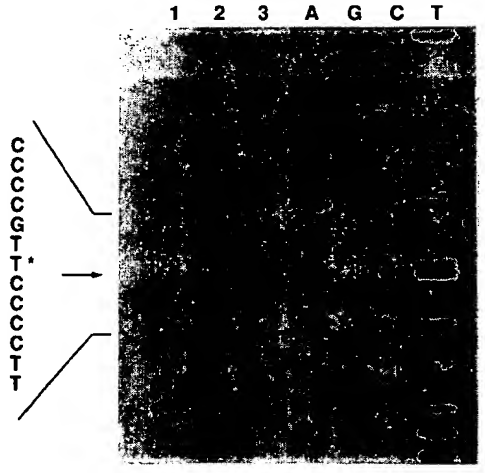

FIG. 5. Mapping of the 5′ end of the *phsA* mRNA. A [γ-^{32}P]ATP end-labeled 24-mer oligonucleotide (5′-GATCTCGGTCTCCCGCGCGTCACC-3′) was annealed to *phsA* mRNA and extended with reverse transcriptase. The reaction products were separated on a sequencing gel with a sequencing ladder, generated by the use of the same primer, to determine the transcription start site of *phsA* mRNA. RNA templates were from *S. lividans* TK24 (lane 1), *S. antibioticus* (lane 2), and *S. lividans* transformed with pIJ2501 (lane 3). The arrow indicates the primer extension product that corresponds to initiation from *phsA*. The sequence on the left is the DNA region around the apparent transcription start site for *phsA*, indicated by an asterisk. Although the band corresponding to the extension product obtained with RNA from *S. antibioticus* is faint in the reproduction shown here (lane 2), it was clearly visible on the original autoradiograms.

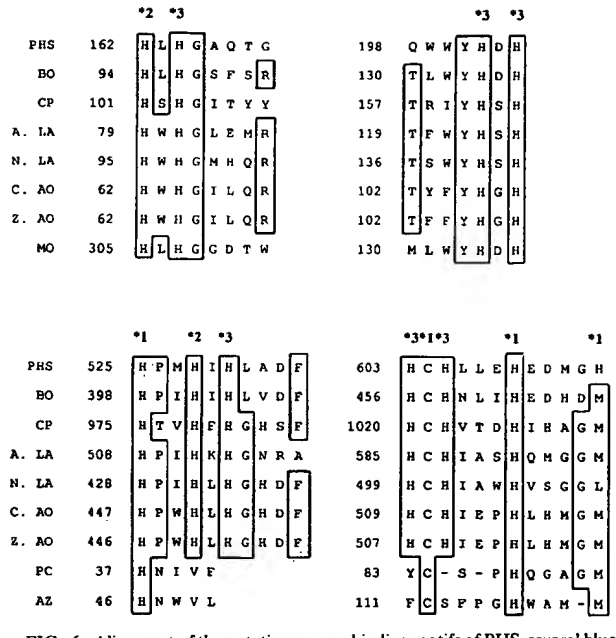

FIG. 6. Alignment of the putative copper-binding motifs of PHS, several blue copper proteins, and manganese-oxidizing protein. Sequence identities between PHS and the other blue copper proteins and manganese-oxidizing protein are boxed. The numbers to the left of the motifs denote the positions in the corresponding protein sequences. The amino acid residues corresponding to potential copper binding sites of three recognized types (40) are shown as follows: type I, *1; type II, *2; type III, *3. Dashes represent gaps introduced to maximize the similarity. Protein sequences were from the following sources: BO, *Myrothecium verrucaria* bilirubin oxidase (32); CP, human ceruloplasmin (33); A. LA, *Aspergillus nidulans* laccase (1); N. LA, *Neurospora crassa* laccase (14); C. AO, cucumber ascorbate oxidase (44); Z. AO, zucchini ascorbate oxidase (40); PC, polar plastocyanin (43); AZ, *Alcaligenes denitrificans* azurin (43); MO, *Leptothrix discophora* manganese-oxidizing protein (8).

(15, 16) than to the −10 consensus sequence, TAGGAT, of *Streptomyces* promoters (48). The −35 region of the putative *phsA* promoter was not strikingly similar to the −35 consensus sequence of either *E. coli* or *Streptomyces* promoters (Fig. 3) (see references 15, 16, and 48). Overall, the *phsA* promoter is not strongly homologous to any promoters for other antibiotic genes from *Streptomyces* spp. (48). However, recent studies do suggest similarities to the P2 promoter of the agarase gene (*dagA*) from *Streptomyces coelicolor* (reference 5 and unpublished data). Preliminary data also suggest that the *phsA* promoter is recognized by an alternative σ factor, σ^E (34). The role of this σ factor in *S. antibioticus* will be described in a subsequent publication.

One noteworthy feature of the *phsA* sequence is the presence of several sets of direct and inverted repeats near the promoter region (Fig. 3). This is especially interesting since some direct and inverted repeat sequences have been reported to be involved in the regulation of gene expression in streptomycetes. For example, repeated sequences have been implicated in the catabolite control of *Streptomyces* genes, including the chitinase genes of *Streptomyces plicatus* (9), the *galP1* promoter of the galactose operon of *S. lividans* (41), and α-amylase promoters of *Streptomyces limosus* (51). None of the *phsA* direct or inverted repeat sequences is strikingly similar to the repeat sequences in the studies described above. However, it is possible that repeat motifs are a common feature of the regions involved in catabolite repression of streptomycete genes. The *phsA* sequence also contains two TNTNAN elements, located within the −10 region and upstream of the *phsA* promoter region (Fig. 3). TNTNAN hexamers were suggested to play a role in *galP1* regulation in *S. lividans* (41).

The predicted amino acid sequence of the PHS subunit resembles that of proteins belonging to the blue copper protein family. Like most members of this group (32), the sequence of the PHS subunit contains four consensus domains (1 to 4) that are presumed to bind the copper ligands (Fig. 6). Domains 1 and 2 are located at the N-terminal portion of the protein, whereas domains 3 and 4 are nearer the C terminus. Even though manganese-oxidizing protein does not belong to the blue copper protein family, similarities were observed in the copper binding domains 1 and 2 between PHS and the manganese-oxidizing protein (Fig. 6). In spite of the diverse distribution of these proteins and their utilization of very different substrates, they all use molecular oxygen in the reactions they catalyze. Although the active sites of these enzymes have not been characterized, the conservation of the copper binding sites strongly suggests their involvement in substrate recognition and catalysis.

In this study, we provided evidence for the regulation of the *phsA* promoter by catabolite repression. Thus, growth of *S. antibioticus* containing the cloned *phsA* promoter on glucose or glucose plus galactose led to a significant inhibition of *xylE* expression from pJSE935 as compared with that of cultures grown on galactose alone (Fig. 7). An identical pattern of inhibition was observed for PHS activity in the same cultures. There are several important implications of this result. First, the data strongly suggest that the sequences required for catabolite repression are contained within the *Bsr*BI fragment cloned in pJSE935. This fragment lacks the repeats 1, 2, and 8 of Fig. 3. Thus, those sequences are presumably not required for catabolite repression. Second, it is clear that whatever the mechanism of catabolite repression in *S. antibioticus*, the relevant machinery can act simultaneously on the endogenous *phs* promoter and on the cloned sequence, since PHS activity parallels *xylE* activity in the experiments illustrated in Fig. 7. With regard to that mechanism, we presented evidence above that it may not involve a simple interaction between an operator and

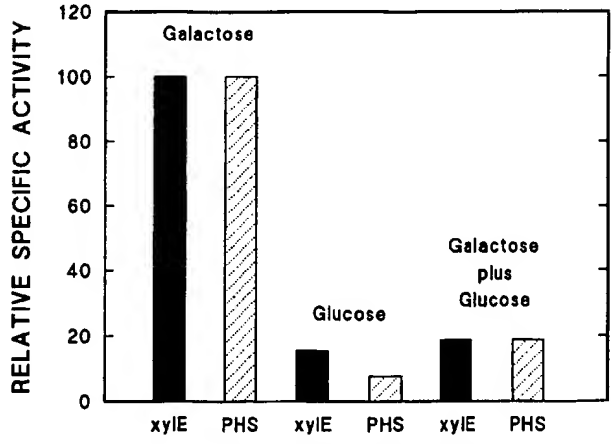

FIG. 7. Effects of carbon source on the expression of the *phsA* promoter in *S. antibioticus*. Transformants containing pJSE935 were grown on galactose, glucose, or glucose plus galactose as described in Materials and Methods. The figures shows the results of catechol dioxygenase and PHS assays of extracts of mycelium harvested 12 h after inoculation. Results represent the averages of three replicates. The values obtained for extracts grown on galactose (42.4 ± 2.1 mU/mg of protein for catechol dioxygenase and 75.6 ± 5.3 U/mg of protein for PHS) were arbitrarily set at 100 for purposes of presentation.

a repressor as it was not possible to release expression of the endogenous *phs* gene from catabolite repression by cloning the disrupted gene at high copy. It is possible, of course, that the repressor binding site involves sequences that were disrupted by the insertion of the *Xba*I linker. It should be possible to distinguish between these possibilities and to learn more about the mechanism of catabolite repression of *phsA* expression by gel mobility shift assays.

ACKNOWLEDGMENTS

We thank Mervyn Bibb for the FRAME analysis.
This work was supported by a grant from Emory University.

REFERENCES

1. **Aramayo, R., and W. E. Timberlake.** 1990. Sequence and molecular structure of the *Aspergillus nidulans yA* (laccase I) gene. Nucleic Acids Res. **18**:3415.
2. **Barry, C. E., P. G. Nayar, and T. P. Begley.** 1989. Phenoxazinone synthase: mechanism for the formation of the phenoxazinone chromophore of actinomycin. Biochemistry **28**:6323–6333.
3. **Bibb, M. J., P. R. Findlay, and M. W. Johnson.** 1984. The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene **30**:157–166.
4. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7**:1513–1523.
5. **Buttner, M. J., A. M. Smith, and M. J. Bibb.** 1988. At least three different RNA polymerase holoenzymes direct transcription of the agarase gene (*dagA*) of *Streptomyces coelicolor* A3(2). Cell **52**:599–607.
6. **Choy, H. A., and G. H. Jones.** 1981. Phenoxazinone synthase from *Streptomyces antibioticus*: purification of the large and small enzyme forms. Arch. Biochem. Biophys. **211**:55–65.
7. **Clayton, T. M., and M. J. Bibb.** 1990. *Streptomyces* promoter-probe plasmids that utilize the *xylE* gene of *Pseudomonas putida*. Nucleic Acids Res. **18**:1077.
8. **Corstjens, P. L. A. M.** 1993. Ph.D thesis. Rijksuniversiteit, Leiden, The Netherlands.
9. **Delic, I., P. Robbins, and J. Westpheling.** 1992. Direct repeat sequences are implicated in the regulation of two *Streptomyces* chitinase promoters that are subject to carbon catabolite control. Proc. Natl. Acad. Sci. USA **89**:1885–1889.
10. **Favre, D.** 1992. Improved phenol-based method for the isolation of DNA fragments from low melting temperature agarose gels. BioTechniques **13**: 24–26.
11. **Fawaz, F., and G. H. Jones.** 1988. Actinomycin synthesis in *Streptomyces antibioticus*: purification and properties of 3-hydroxyanthranilate 4-methyltransferase. J. Biol. Chem. **263**:4602–4606.
12. **Freeman, J. C., P. G. Nayar, T. P. Begley, and J. J. Villafranca.** 1993. Stoichiometry and spectroscopic identity of copper centers in phenoxazinone synthase: a new addition to the blue copper oxidase family. Biochemistry **32**:4826–4830.
13. **Gallo, M., and E. Katz.** 1972. Regulation of secondary metabolite biosynthesis: catabolite repression of phenoxazinone synthase and actinomycin formation by glucose. J. Bacteriol. **109**:659–667.
14. **Germann, U. A., G. Muller, P. E. Hunziker, and K. Lerch.** 1988. Characterization of two allelic forms of *Neurospora crassa* laccase. J. Biol. Chem. **263**:885–896.
15. **Harley, C. B., and R. P. Reynolds.** 1987. Analysis of *E. coli* promoter sequences. Nucleic Acids Res. **15**:2343–2361.
16. **Hawley, D. K., and W. R. McClure.** 1983. Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Res. **11**:2237–2255.
17. **Hopwood, D. A., M. J. Bibb, K. F. Chater, T. Kieser, C. J. Bruton, H. M. Kieser, D. J. Lydiate, C. P. Smith, J. M. Ward, and H. Schrempf.** 1985. Genetic manipulation of streptomyces: a laboratory manual. The John Innes Foundation, Norwich, United Kingdom.
18. **Hopwood, D. A., T. Kieser, H. M. Wright, and M. J. Bibb.** 1983. Plasmids, recombination and chromosome mapping in Streptomyces lividans. J. Gen. Microbiol. **129**:2257–2269.
19. **Ingram, C., M. Brawner, P. Youngman, and J. Westpheling.** 1989. *xylE* functions as an efficient reporter gene in *Streptomyces* spp.: use for the study of *galP1*, a catabolite-controlled promoter. J. Bacteriol. **171**:6617–6624.
20. **Jones, G. H.** 1985. Regulation of phenoxazinone synthase expression in *Streptomyces antibioticus*. J. Bacteriol. **163**:1215–1221.
21. **Jones, G. H.** 1985. Regulation of actinomycin synthesis in *Streptomyces antibioticus*. J. Nat. Prod. (Lloydia) **49**:981–987.
22. **Jones, G. H.** 1987. Actinomycin synthesis in *Streptomyces antibioticus*: enzymatic conversion of 3-hydroxyanthranilic acid to 4-methyl-3-hydroxyanthranilic acid. J. Bacteriol. **169**:5575–5578.
23. **Jones, G. H.** 1993. Combined purification of actinomycin synthetase I and 3-hydroxyanthranilic acid 4-methyltransferase from *Streptomyces antibioticus*. J. Biol. Chem. **268**:6831–6834.
24. **Jones, G. H., and D. A. Hopwood.** 1984. Activation of phenoxazinone synthase expression in *Streptomyces lividans* by cloned DNA sequences from *Streptomyces antibioticus*. J. Biol. Chem. **259**:14158–14164.
25. **Jones, G. H., and D. A. Hopwood.** 1984. Molecular cloning and expression of the phenoxazinone synthase gene from *Streptomyces antibioticus*. J. Biol. Chem. **259**:14151–14157.
26. **Katayama, C.** 1990. Single-stranded DNA sequencing: a simplified protocol for single-stranded DNA rescue. Strategies **4**:56–58.
27. **Katz, E., C. J. Thompson, and D. A. Hopwood.** 1983. Cloning and expression of the tyrosinase gene from *Streptomyces antibioticus* in *Streptomyces lividans*. J. Gen. Microbiol. **129**:2703–2714.
28. **Katz, E., and H. Weissbach.** 1962. Biosynthesis of the actinomycin chromophore; enzymatic conversion of 4-methyl-3-hydroxyanthranilic acid to actinocin. J. Biol. Chem. **237**:882–886.
29. **Keller, U.** 1987. Actinomycin synthetases; multifunctional enzymes responsible for the synthesis of the peptide chains of actinomycin. J. Biol. Chem. **262**:5852–5856.
30. **Keller, U., H. Kleinkauf, and R. Zocher.** 1984. 4-methyl 3-hydroxyanthranilic acid activating enzyme from actinomycin producing *Streptomyces chrysomallus*. Biochemistry **23**:1479–1484.
31. **Keller, U., and W. Schlumbohm.** 1992. Purification and characterization of actinomycin synthetase I, a 4-methyl-3-hydroxyanthranilic acid-AMP ligase from *Streptomyces chrysomallus*. J. Biol. Chem. **267**:11745–11752.
32. **Koikeda, S., K. Ando, H. Kaji, T. Inoue, S. Murao, K. Takeuchi, and T. Samejima.** 1993. Molecular cloning of the gene for bilirubin oxidase from *Myrothecium verrucaria* and its expression in yeast. J. Biol. Chem. **268**:18801–18809.
33. **Koschinsky, M. L., W. D. Funk, B. A. v. Oost, and R. T. A. MacGillivray.** 1986. Complete cDNA sequence of human preceruloplasmin. Proc. Natl. Acad. Sci. USA **83**:5086–5090.
34. **Lonetto, M., K. L. Brown, K. Rudd, and M. J. Buttner.** 1994. Analysis of the *Streptomyces coelicolor sigE* gene reveals a new sub-family of eubacterial RNA polymerase σ factors involved in the regulation of extracytoplasmic functions. Proc. Natl. Acad. Sci. USA **91**:7573–7577.
35. **Luria, S. E., J. N. Adams, and R. C. Ting.** 1960. Transduction of lactose utilizing ability among strains of *E. coli* and *S. dysenteriae* and the properties of the transducing phage particles. Virology **12**:348–356.
36. **Lydiate, D. J., F. Malpartida, and D. A. Hopwood.** 1985. The *Streptomyces* plasmid SCP2*: its functional analysis and development into useful cloning vectors. Gene **35**:223–235.
37. **Madu, A. C., and G. H. Jones.** 1989. Molecular cloning and *in vitro* expression of a silent phenoxazinone synthase gene from *Streptomyces lividans*. Gene **84**:287–294.
38. **Malpartida, F., and D. A. Hopwood.** 1984. Molecular cloning of the whole biosynthetic pathway of a *Streptomyces* antibiotic and its expression in a heterologous host. Nature (London) **309**:462–464.
39. **Marshall, R., B. G. Redfield, E. Katz, and H. Weissbach.** 1968. Changes in phenoxazinone synthase activity during the growth cycle of *Streptomyces antibioticus*. Arch. Biochem. Biophys. **123**:317–323.
40. **Masserschmidt, A., A. Rossi, R. Ladenstein, R. Huber, M. Bolognesi, G. Gatti, A. Marchesini, R. Petruzzelli, and A. Finazzi-Agro.** 1989. X-ray crystal structure of the blue oxidase ascorbate oxidase from zucchini: analysis of the polypeptide fold and a model of the copper sites and ligands. J. Mol. Biol. **206**:513–529.
41. **Mattern, S. G., M. E. Brawner, and J. Westpheling.** 1993. Identification of a complex operator for *galP1*, the glucose-sensitive, galactose-dependent promoter of the *Streptomyces* galactose operon. J. Bacteriol. **175**:1213–1220.
42. **Moran, C. P., Jr.** 1990. Measuring gene expression in *Bacillus*, p. 267–293. *In* C. R. Harwood and S. M. Cutting (ed.), Molecular biology methods for *Bacillus*. John Wiley & Sons, Inc., New York.
43. **Norris, G. E., B. F. Anderson, and E. N. Baker.** 1983. Structure of azurin from *Alcaligenes denitrificans* at 2.5 A resolution. J. Mol. Biol. **165**:501–521.
44. **Ohkawa, J., N. Okada, A. Shinmyo, and M. Takano.** 1989. Primary structure of cucumber (*Cucumis sativus*) ascorbate oxidase deduced from cDNA sequence: homology with blue copper proteins and tissue-specific expression. Proc. Natl. Acad. Sci. USA **86**:1239–1243.
45. **Sala-Trepat, J. M., and W. C. Evans.** 1971. The meta cleavage of catechol by *Azotobacter* species 4-oxalocrotonate pathway. Eur. J. Biochem. **20**:400–413.
46. **Sambrook, J., E. F. Fritsch, and T. E. Maniatis.** 1989. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
47. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
48. **Strohl, W. R.** 1992. Compilation and analysis of DNA sequences associated with apparent streptomycete promoter. Nucleic Acids Res. **20**:961–974.
49. **Thompson, C. J., J. M. Ward, and D. A. Hopwood.** 1982. Cloning of antibiotic resistance and nutritional genes in streptomycetes. J. Bacteriol. **151**:668–677.
50. **Troost, T., and E. Katz.** 1979. Phenoxazinone biosynthesis: accumulation of

a precursor, 4-methyl-3-hydroxyanthranilic acid, by mutants of *Streptomyces parvulus*. J. Bacteriol. **169**:5575–5578.
51. **Virolle, M.-J., and J. Gagnat.** 1994. Sequences involved in growth-phase-dependent expression and glucose repression of a *Streptomyces* α-amylase gene. Microbiology **140**:1059–1067.
52. **Waksman, S. A.** 1968. Actinomycin: nature, formation and activities. John Wiley & Sons, Inc., New York.
53. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33**:103–119.
54. **Zukowski, M. M., D. F. Gaffney, D. Speck, M. Kauffmann, A. Findeli, A. Wisecup, and J.-P. Lecocq.** 1983. Chromogenic identification of genetic regulatory signals in *Bacillus subtilis* based on expression of a cloned *Pseudomonas* gene. Proc. Natl. Acad. Sci. USA **80**:1101–1105.

JOURNAL OF VIROLOGY, Dec. 1993, p. 7705–7710
0022-538X/93/127705-06$02.00/0

Vol. 67, No. 12

Hierarchy of Polyadenylation Site Usage by Bovine Papillomavirus in Transformed Mouse Cells

ELLEN M. ANDREWS AND DANIEL DiMAIO*

Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510-8005

Received 4 August 1993/Accepted 16 September 1993

The great majority of viral mRNAs in mouse C127 cells transformed by bovine papillomavirus type 1 (BPV) have a common 3′ end at the early polyadenylation site which is 23 nucleotides (nt) downstream of a canonical poly(A) consensus signal. Twenty percent of BPV mRNA from productively infected cells bypasses the early polyadenylation site and uses the late polyadenylation site approximately 3,000 nt downstream. To inactivate the BPV early polyadenylation site, the early poly(A) consensus signal was mutated from AAUAAA to UGUAAA. Surprisingly, this mutation did not result in significant read-through expression of downstream RNA. Rather, RNA mapping and cDNA cloning experiments demonstrate that virtually all of the mutant RNA is cleaved and polyadenylated at heterogeneous sites approximately 100 nt upstream of the wild-type early polyadenylation site. In addition, cells transformed by wild-type BPV harbor a small population of mRNAs with 3′ ends located in this upstream region. These experiments demonstrate that inactivation of the major poly(A) signal induces preferential use of otherwise very minor upstream poly(A) sites. Mutational analysis suggests that polyadenylation at the minor sites is controlled, at least in part, by UAUAUA, an unusual variant of the poly(A) consensus signal approximately 25 nt upstream of the minor polyadenylation sites. These experiments indicate that inactivation of the major early polyadenylation signal is not sufficient to induce expression of the BPV late genes in transformed mouse cells.

Eukaryotic RNA polymerase II transcription units are typically transcribed past the mature mRNA 3′ end. These transcripts are then cleaved and a poly(A) tract of 200 to 300 nucleotides (nt) is added to generate the 3′ end of the mature mRNA (for reviews, see references 30 and 37). Eighty to ninety percent of animal cell mRNAs contain the sequence AAUAAA 10 to 30 nt upstream of the poly(A) tail. Another 10% have the variant AUUAAA; other variants are rare (38). These consensus sequences have been shown to be required for efficient and accurate cleavage and polyadenylation both in vivo and in vitro (17, 27, 29). Generally, when this sequence is mutated, polyadenylation occurs at a downstream site, often with reduced efficiency (17). The region upstream and downstream of the AAUAAA consensus signal, including GU-rich downstream sequences, has also been identified as playing a role in the cleavage and polyadenylation of some transcripts (30, 37).

Bovine papillomavirus type 1 (BPV) induces fibropapillomas in cattle and transforms a number of cultured rodent fibroblast cell lines to tumorigenicity. The papillomaviruses are unable to propagate in such transformed cells, in part because the early polyadenylation site used by essentially all BPV transcripts in transformed cells is located between the transcriptional promoters and L1 and L2, the two genes which encode the virion proteins (Fig. 1A) (16, 23, 39). Similarly, in BPV-induced skin fibropapillomas, usage of this early polyadenylation site precludes expression of the capsid protein genes in transformed dermal fibroblasts and presumably in the basal keratinocytes as well (4, 5, 35). In terminally differentiating keratinocytes which express the capsid proteins and produce virus, about 20% of the viral mRNA reads through the early polyadenylation site and is instead polyadenylated approximately 3,000 nt downstream at the late polyadenylation site (5). Thus, regulation of polyadenylation at the early site appears to be crucial for viral late gene expression.

To study signals that control polyadenylation in BPV-transformed mouse C127 cells, we mutated the early poly(A) consensus signal AAUAAA, located 23 nt upstream of the early poly(A) site. It was expected that mutant transcripts would now bypass the early poly(A) site and that late region sequences would be included in stable RNA. RNA mapping experiments instead demonstrated that mutant transcripts were polyadenylated at heterogeneous sites approximately 100 nt upstream of the early polyadenylation site used in cells transformed by wild-type BPV. Evidence is presented which suggests that an unusual variant of the poly(A) consensus sequence, UAUAUA, plays a role in the regulation of polyadenylation at the upstream polyadenylation sites.

Construction and preliminary characterization of the poly(A) consensus mutant. To disrupt polyadenylation at the BPV major early polyadenylation site at nt 4203, oligonucleotide-directed mutagenesis was used to mutate the poly(A) consensus signal at nt 4180 from AAUAAA to UGUAAA (Fig. 1B), thereby creating a new *Pvu*II cleavage site (23a). The resulting mutation on a *Bst*XI-to-*Sal*I fragment was reconstructed into the full-length wild-type BPV genome (clone pBPV-142-6 [33]) to generate mutant pBPV-EPA1. Nucleotide sequence analysis of the fragment replaced in generating pBPV-EPA1 (nt 3849 and 4450) demonstrated that no extraneous mutations were introduced during mutagenesis.

The ability of three isolates of pBPV-EPA1 to transform C127 cells was assayed by determining the efficiency of focus formation after *Bam*HI digestion to release the viral DNA from the plasmid vector and transfection as described previously (14). All three mutant isolates transformed cells with approximately the same efficiency as wild-type BPV DNA (data not shown). Cell lines were derived from pools of foci induced by pBPV-EPA1 (EPA1p) and by wild-type BPV DNA (142-6p). ID13 cells, a C127 cell line transformed by infection with BPV, were used as an additional wild-type control.

* Corresponding author.

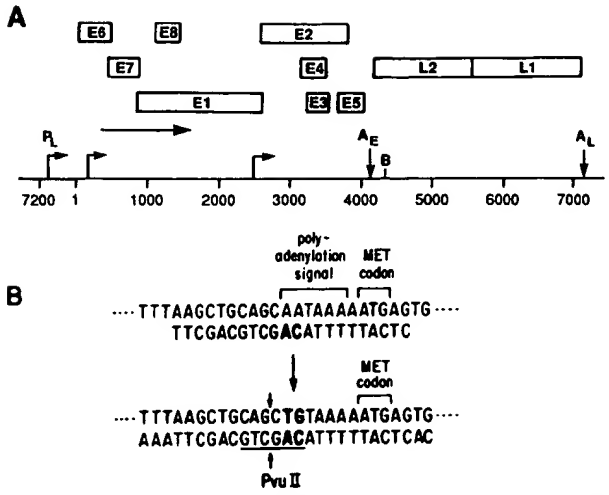

FIG. 1. The BPV genome and design of the EPA1 mutation. (A) The BPV genome linearized at the 3′ end of the late transcription unit is shown. Open boxes indicate translational open reading frames. The long horizontal arrow indicates the direction of transcription. The short horizontal arrows indicate the positions of major promoters. The late promoter is designated P_L. A_E and A_L denote the early and late polyadenylation sites, respectively. B indicates the position of the unique *Bam*HI site. Nucleotide numbers are shown at the bottom. (B) The EPA1 mutation. The top line shows the wild-type BPV DNA sequence around the early polyadenylation signal. The sequence of the mutagenic oligonucleotide L1 is shown directly below it, with the base substitutions shown in boldface. The template was the small *Bam*HI to *Eco*RI fragment of BPV-1 DNA cloned in M13mp8 (13). The sequence at the bottom shows the mutation with the new *Pvu*II site indicated. The open reading frame L2 initiation codon is designated the MET codon.

Southern blot analysis of viral DNA from transformed cells demonstrated that the mutant viral DNA was maintained in transformed cells as a multicopy plasmid without gross rearrangement and with restoration of the *Bam*HI site used to excise the viral DNA from the plasmid vector (data not shown).

Mapping the 3′ end of the mutant mRNA. The mutation in pBPV-EPA1 was designed to eliminate polyadenylation at the wild-type early polyadenylation site immediately upstream of the late open reading frames. Extensive Northern (RNA) blot analysis and RNA protection experiments failed to detect significant amounts of RNA extending past the polyadenylation site into the late region, but these experiments did not exclude the presence of low levels of read-through RNA (data not shown). There was severalfold more stable viral RNA in cells transformed by wild-type BPV than in those transformed by the polyadenylation site mutant (Fig. 2).

RNase protection experiments were performed to map the 3′ ends of the mutant transcripts. ID13 and EPA1p RNAs were assayed for protection of an antisense EPA1 RNA probe spanning the early polyadenylation site at nt 4203 (Fig. 2, left panel). The size of the fragment protected by RNA from ID13 cells indicates that, as expected, the wild-type viral RNA extends past nt 4180, the site of the mutation in the probe (lane c). In contrast, EPA1p RNA protected several fragments approximately 100 nt shorter than those protected by wild-type RNA, suggesting that the mutant RNA is polyadenylated upstream of the normal position (lanes a and b). The difference in the pattern of protected bands between the two EPA1p lanes, a and b, is due to the different cleavage specificities of the two RNases used in these reactions. The same result was obtained with oligo(dT)-selected EPA1p RNA (data not shown), indicating that these shorter species are polyadenylated, a conclusion confirmed by cDNA cloning (see below). There was no evidence of significant polyadenylation of EPA1p RNA at the usual position, nor were prominent shorter novel bands protected in the ID13 sample. RNA from two additional cell lines generated with the original isolate of the mutant and two additional cell lines generated with independent isolates of the mutant showed the protection pattern characteristic of the mutant (data not shown). These results suggested that sequences downstream of nt 4100 were absent from mutant RNA, an interpretation supported by the results of protection experiments with additional antisense probes and the results of Northern blot hybridization experiments with oligonucleotide probes (data not shown). These results are interpreted in the right panel of Fig. 2.

cDNA cloning and sequencing. The results presented above suggest that new heterogeneous polyadenylation sites near nt 4100 are utilized in EPA1p RNA. To confirm this interpretation, the 3′ ends of both wild-type and mutant RNAs were cloned and sequenced. Oligod(T)-selected (3) 142-6p and EPA1p RNAs were reverse transcribed with oligo(dT) as primer and the reagents and protocol of a cDNA synthesis kit (Amersham). The resulting first-strand cDNAs were amplified by the polymerase chain reaction (PCR) method with the primers diagrammed in Fig. 3A (18, 32, 34). To specifically amplify BPV sequences, the upstream PCR primer PCR5 corresponded to BPV nt 3998 to 4031. To selectively amplify polyadenylated molecules, the downstream PCR primer PCRT was 5′ d(GGGGATCCT$_{25}$) 3′, which hybridized to any product containing a poly(A) tract. Annealing was carried out at 25°C, because PCRT has a calculated T_m of 38.9°C in PCR buffer conditions (32). The products of each amplification reaction were cloned into pUC18, and colonies containing an insert were identified by colony hybridization (22) with an oligonucleotide probe PCR1 complementary to a region (nt 4063 to 4089) between the upstream primer and the proposed 3′ end of mutant RNA.

The results of sequence analysis of the cDNA clones are summarized in Fig. 3B. Sites of polyadenylation were identified as junctions between BPV DNA sequence and tracts of poly(A). Six of the 11 clones derived from cells transformed by wild-type BPV were polyadenylated after nt 4203, the previously described early polyadenylation site (39), thus validating this strategy of identifying polyadenylation sites. In contrast, none of the clones derived from mutant RNA displayed the wild-type polyadenylation site. Instead, seven of the nine EPA1p clones contain a stretch of poly(A) immediately after BPV nt 4107, and the other two clones contain poly(A) after nt 4101 and 4092. These results are consistent with the RNase protection and Northern blot results which indicate the existence of heterogeneous 3′ ends near nt 4100 in mutant RNA and demonstrate that these new 3′ ends are in fact new sites of polyadenylation. Interestingly, the anomalous clones (almost half) derived from wild-type RNA showed polyadenylation at heterogeneous sites similar to those found with the mutant RNA. These results indicate that there is a population of mRNAs with heterogeneous polyadenylation sites around nt 4100 in cells transformed by wild-type BPV. The preferential amplification of shorter PCR products may explain the relatively frequent isolation of these shorter cDNAs from cells transformed by wild-type BPV. We have occasionally observed faint bands in protection experiments with ID13 RNA which

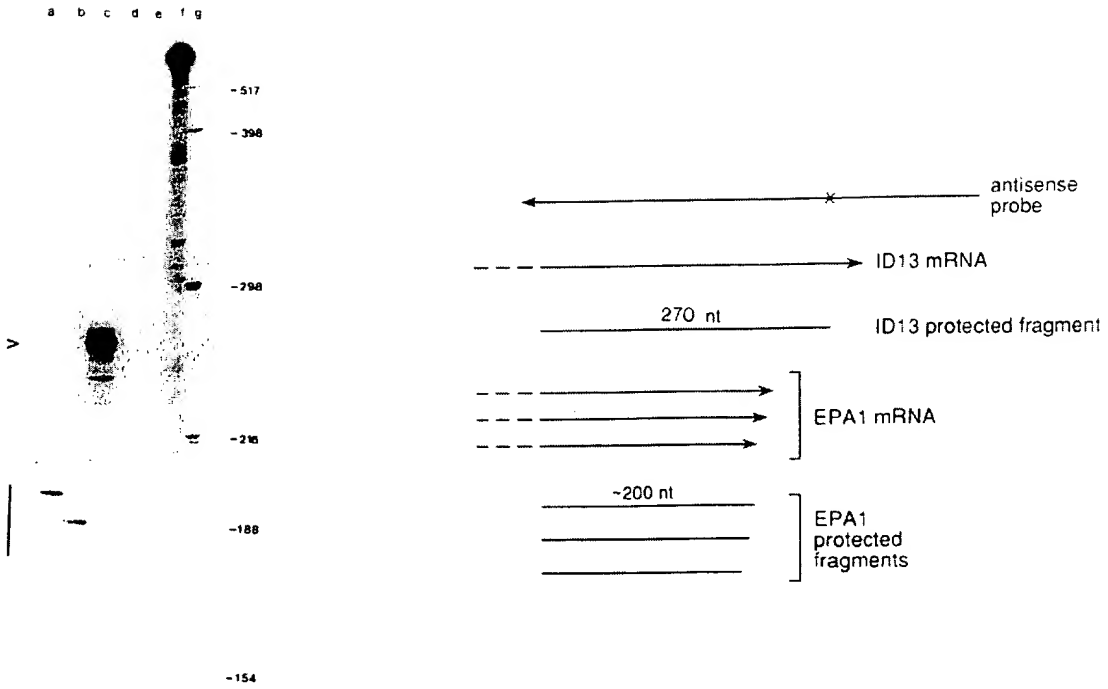

FIG. 2. RNase protection analysis of viral early region RNA in transformed cells. (Left panel) Ten micrograms of total cellular RNA (9, 21) from EPA1p (lanes a and b), ID13 (lane c), and C127 (lane d) cells was hybridized to an antisense EPA1 RNA probe (28) complementary to BPV nt 3912 to 4450, spanning the position of the normal early polyadenylation site but containing the mutation at the polyadenylation signal. Hybrids were digested with either 2 μg of RNase T1 per ml (lane a), 40 μg of RNase A per ml (lane b), or a mixture of both RNases (lanes c to e), and protected fragments were detected by autoradiography after electrophoresis through a 4% polyacrylamide–50% urea gel. The sample in lane e was the probe digested after mock hybridization; the sample in lane f is undigested probe. The nucleotide lengths of size markers in lane g are indicated. The arrowhead indicates the predicted position of a 270-nt fragment extending from the 3' end of the probe to the site of the mutation, which is generated by cleavage at the mismatch between the wild-type RNA and the mutation in the probe. The vertical line on the left indicates the small cluster of bands protected by mutant RNA. (Right panel) Schematic representation of the probe, protected fragments generated by RNase digestion, and the deduced structure of viral RNA species. Arrows indicate the direction of transcription, with the arrowheads representing the 3' end of each transcript. The X indicates the position of the mutation in the probe.

are consistent with minor sites of polyadenylation at these upstream positions (data not shown).

Identification of a signal controlling polyadenylation at the upstream sites. There is no poly(A) consensus sequence or previously described functional variant within 100 bp upstream of nt 4100. However, the sequence UAUAUA is present at nt 4073, approximately 30 nt 5' to the poly(A) sites in EPA1p RNA (Fig. 3B). It is the closest match to the consensus sequence in the region, and it appears to be in the appropriate position to specify cleavage at the sites detected in mutant RNA. To test the role of this sequence in specifying polyadenylation in the absence of the wild-type signal, it was mutated from UAUAUA to GAUAUC by using the mutagenic primer 5' d(AACTTCATAC AGGATATCAA ACAAATCA)3', corresponding to BPV sequence from nt 4063 to 4090, and single-stranded EPA1 DNA as a template. The resulting mutant, pBPV-EPA2 (see Fig. 5) therefore contained both the original mutation at the poly(A) consensus signal and the new mutations in the putative variant signal. This mutant transformed C127 cells with approximately wild-type efficiency, and RNA from a pooled cell line transformed by EPA2 DNA was mapped by using RNase protection and an antisense EPA2 probe (Fig. 4). RNA from ID13 cells protected the fragment sizes predicted if cleavage occurred at the sites of mismatch between wild-type RNA and the probe (which contains mutations at nt 4073, 4078, 4180, and 4181) (lane b). EPA2 RNA protected two major size classes of fragments (lane a). One was a set of probe fragments approximately 190 to 200 nt long, corresponding to polyadenylation near nt 4100 as in EPA1p RNA. These protected fragments comigrate with the fragments protected by EPA1 RNA (data not shown) and are the size predicted if polyadenylation occurred at nt 4107, the mutant site mapped by cDNA cloning. In addition, EPA2p RNA protected several longer fragments corresponding to heterogeneous RNA 3' ends between nt 4200 and 4450. The EPA2 mutation thus reduced the efficiency with which the upstream polyadenylation sites are used, but it did not appear to affect the position of poly(A) addition for those transcripts that are successfully polyadenylated in this region. These results indicate that the UAUAUA plays a role in specifying the new upstream sites of polyadenylation in EPA1p RNA.

Discussion. These experiments were designed to study polyadenylation site usage in BPV-transformed mouse cells. A point mutation in the poly(A) consensus signal disrupted polyadenylation at that site both in vivo, as demonstrated here, and in an in vitro polyadenylation system (24). RNase protection, Northern blotting, and cDNA cloning and sequencing established that stable mutant transcripts utilized heteroge-

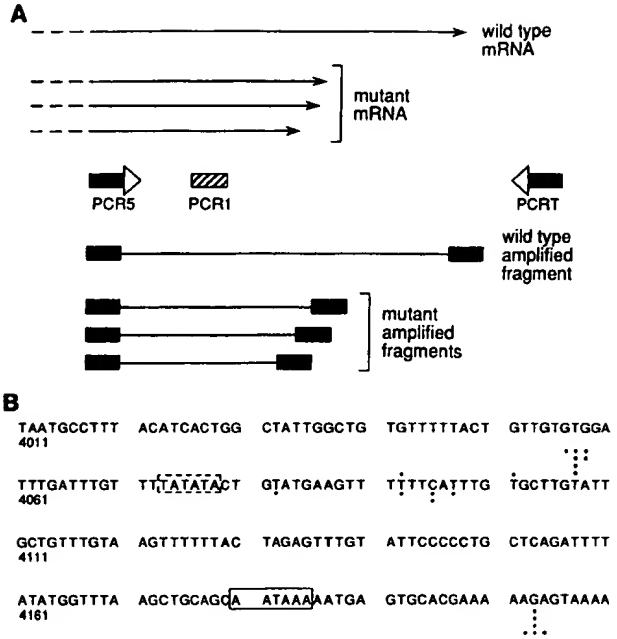

FIG. 3. (A) PCR-based strategy to clone the 3′ ends of viral RNA. The top portion of the panel shows the deduced structure of the 3′ ends of the major viral RNAs. The upstream primer, PCR5, is complementary to all known wild-type and mutant viral early RNAs. The downstream primer, PCRT, consists of oligo(dT) and a cloning site but no BPV-specific sequences. After amplification with cDNA reverse transcribed from polyadenylated RNA as a template, BPV cDNAs were cloned into pUC18, identified by hybridization to PCR1, and sequenced. (B) cDNA clone sequences. The sense strand BPV sequence from nt 4011 to 4210 is shown. The normal poly(A) consensus signal is enclosed in the solid line, and the putative upstream poly(A) signal is enclosed in the dashed line. Each dot indicates the position of a junction between BPV DNA and the poly(A) tract in a cDNA clone. Clones derived by amplification of wild-type RNA are represented by dots below the sequence, and those derived from mutant RNA are represented by dots above the sequence.

neous polyadenylation sites approximately 100 nt upstream of the wild-type polyadenylation site. The results of the cDNA cloning also demonstrated that some wild-type transcripts have heterogeneous 3′ ends in the region used by the mutant RNAs, indicating that this is a minor polyadenylation site in cells transformed by wild-type BPV. In a related system, Doniger et al. (15) found usage of upstream polyadenylation sites by human papillomavirus type 16 transcripts in an immortalized human exocervical epithelial cell line harboring a human papillomavirus type 16 genome with an extensive deletion immediately downstream of a wild-type early poly(A) signal. The 3′ ends of viral RNA from these cells mapped to both the normal site and to a heterogeneous region 400 to 500 nt upstream of that site.

The results described here suggest a hierarchy of polyadenylation site usage in BPV-transformed cells, as is summarized in Fig. 5. Wild-type BPV mRNA is polyadenylated at the major polyadenylation site at nt 4203, with a small fraction of transcripts being polyadenylated at minor upstream sites around nt 4100. When the major signal is disrupted (as in EPA1), the sites around nt 4100 become the predominant sites of polyadenylation. When the major signal is inactivated and the minor signal is partially disrupted (as in EPA2), both the upstream sites and new downstream sites between nt 4200 and 4450 are used. Additional experiments have shown that polyadenylation occurs exclusively at these downstream sites when the major signal is inactivated and the upstream polyadenylation region is deleted (2). There are several potential polyadenylation signals in this downstream region, including a sequence at nt 4304 that deviates by 1 nt from the consensus polyadenylation signal. In addition, Burnett et al. (8) observed polyadenylation near nt 4450 in RNA from cells transformed by a spontaneous BPV-1 deletion mutant lacking the major poly(A) site and surrounding sequences. One can speculate that the function of the multiple potential early polyadenylation sites in BPV is to ensure that late genes are not expressed under inappropriate conditions, for example in transformed dermal fibroblasts or basal epidermal keratinocytes.

Polyadenylation site selection appears to be a complex process that takes into account both the relative strengths of potential sites and their positions relative to one another (12, 20). Moreover, the representation of polyadenylation sites in stable RNA reflects a number of factors in addition to polyadenylation site selection, including the stability of various RNA species. The results of the RNase protection experiments reported here indicate that the upstream polyadenylation sites are used far more abundantly by the early polyadenylation signal mutant than by the wild type. However, it is also clear that there is less total viral RNA in cells transformed by the mutant. It is possible that processing at the upstream sites remains relatively inefficient even with the mutant polyadenylation signal, resulting in the synthesis of a rapidly degraded pool of unprocessed RNA extending into the late region. In fact, Furth and Baker (19) have described a sequence element in the BPV late region which prevents the accumulation of stable viral RNA in transformed cells.

The closest match to a poly(A) consensus signal in the vicinity of the minor upstream polyadenylation sites is UAUAUA, approximately 25 nt upstream of the new RNA 3′ ends. RNA from cells containing mutations of both the original poly(A) signal and this putative upstream signal contains heterogeneous 3′ ends at both the upstream sites and at additional positions downstream of the normal site. This result suggests that the UAUAUA plays a role in directing polyadenylation at the upstream polyadenylation sites and that the mutation did not fully disrupt the function of the UAUAUA sequence. We are not aware of a precedent for UAUAUA acting as a poly(A) signal in mammalian cells, although it can direct mRNA 3′ end formation and polyadenylation in *Saccharomyes cerevisiae* (31). However, we note that the region around the upstream cleavage sites contains numerous oligo(dT) tracks and GT dinucleotides, sequence motifs found near some bona fide mammalian poly(A) signals.

The wild-type poly(A) consensus signal appears to suppress utilization of the variant signal located approximately 100 nt upstream. Such suppression may be rather general. Connelly and Manley (10) studied the simian virus 40 early polyadenylation region, which contains two closely spaced AAUAAA signals. In the wild-type situation, only the 3′ site is efficiently utilized. However, if this preferred site was inactivated by mutation, increased usage of the 5′ site was observed. In addition, Denome and Cole (11) showed that addition of tandemly arranged polyadenylation signals decreased usage of the upstream site. These findings imply that genomes may contain numerous potential sites of polyadenylation whose activity is suppressed by the relatively close apposition of another polyadenylation signal, which perhaps competes more efficiently for a limiting polyadenylation factor. Therefore, alternative polyadenylation, which is a well-documented con-

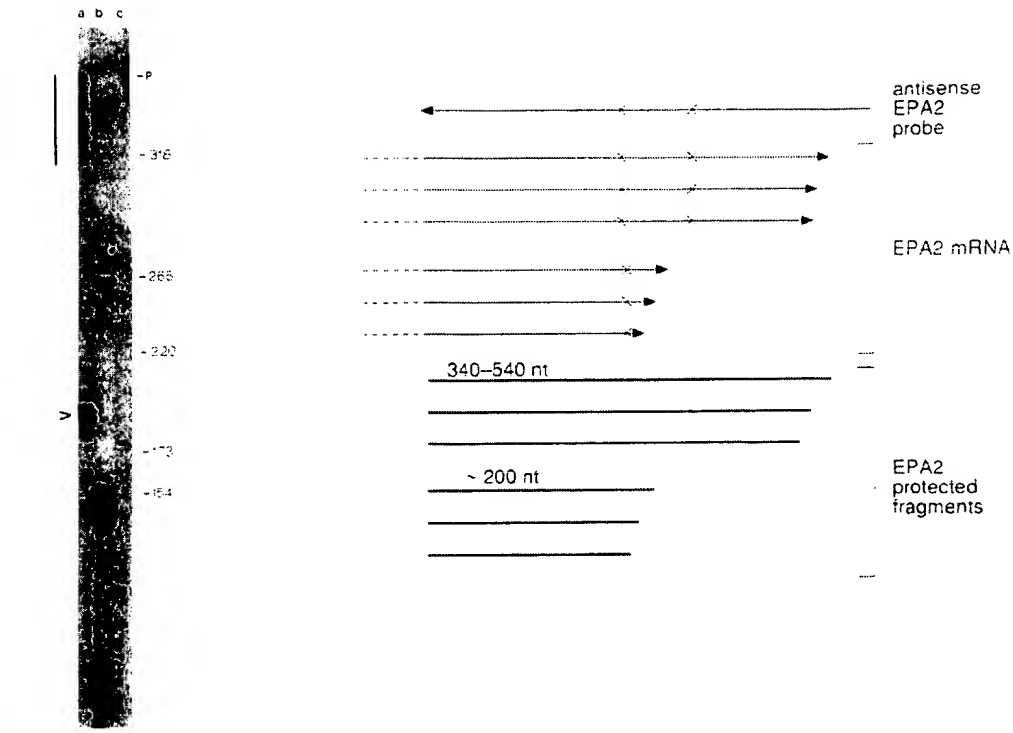

FIG. 4. Evidence that UAUAUA at nt 4073 plays a role in poly(A) site selection. (Left panel) Radiolabelled antisense RNA probe extending from BPV nt 4450 to 3912 was transcribed in vitro from EPA2, hybridized to 10 μg of cellular RNA isolated from EPA2p (lane a), ID13 (lane b), or C127 (lane c) cells, and digested with a mixture of RNases A and T1. Protected fragments were subjected to polyacrylamide gel electrophoresis and detected by autoradiography. The arrowhead on the left indicates the position of approximately 190- to 200-nt probe fragments extending from the 3′ end of the probe to the upstream sites of polyadenylation around BPV nt 4100. The vertical line on the left indicates the position of the larger fragments also protected by mutant RNA. The approximately 265-base fragment in lane b appears to be derived from partially digested hybrids. The lengths (in nucleotides) of coelectrophoresed size markers are shown. P indicates the position of undigested probe (538 nt). (Right panel) Schematic representation of the antisense probe, sizes of the protected fragments, and deduced structures of viral RNA species. The X's show the positions of the mutations at the upstream and downstream polyadenylation signals in the probe and in RNA isolated from cells transformed by the double mutant.

trol point for regulating gene expression (25), may result in some cases from inactivation of a preferred polyadenylation signal rather than by direct activation of a suboptimal one.

The mechanism by which BPV prevents expression of viral

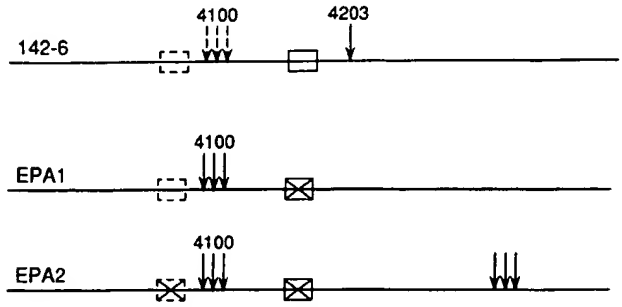

FIG. 5. Usage of early region polyadenylation sites. The horizontal lines represent the region of the BPV genome around the early polyadenylation site for wild-type BPV DNA (142-6) and the indicated mutants. Transcription proceeds from left to right. The unbroken box represents the normal early polyadenylation consensus signal at nt 4180, and the dashed boxes represent the putative upstream polyadenylation signal at nt 4073. The vertical arrows indicate the positions of poly(A) addition, and triple arrows indicate heterogeneous polyadenylation sites, with minor sites represented as dashed arrows. Boxes containing an X indicate a mutant polyadenylation signal.

late genes in transformed cells but allows their expression in differentiated keratinocytes is central to an understanding of papillomavirus biology. One level of restriction in transformed cells is clearly at the level of stable mRNA accumulation, because little or no BPV mRNA from the late region is present in cultured fibroblasts. Analysis of nascent RNA from ID13 cells indicates that at least 90% of BPV transcripts terminate between the early and late poly(A) sites and therefore never reach the late poly(A) signal (6). The mechanism(s) allowing production of late RNAs during natural infection may act primarily at the level of polyadenylation site selection, or it may act at some other steps in mRNA biogenesis, such as alterations in promoter usage, splicing patterns, or transcription termination, which secondarily affect cleavage and polyadenylation (for examples, see references 1, 7, 26, and 36). However, the results presented here indicate that specific suppression of the major early polyadenylation signal is unlikely to be the sole step in releasing the block to BPV late gene expression, because inhibition of polyadenylation at additional potential early sites must also occur. The mechanism involved in late gene expression must coordinately suppress cleavage and polyadenylation at multiple potential sites near the 3′ end of the early region in some of the transcripts, while many transcripts are still polyadenylated at the early polyadenylation site. Regulation of BPV late gene expression is clearly a complex process and bypass of the early major

polyadenylation site is a necessary but not sufficient component of that process. The study of BPV transcriptional regulation promises to provide insights into not only papillomavirus biology but also the mechanisms of regulation of gene expression in general.

We thank S. Lee, S. Amara, and D. Ward for helpful discussions; S. Baserga and J. Brandsma for reviewing the manuscript; and J. Zulkeski for assistance in manuscript preparation.

This work was supported by a grant from the National Institutes of Health (CA37157).

REFERENCES

1. **Acheson, N. H.** 1984. Kinetics and efficiency of polyadenylation of late polyomavirus nuclear RNA: generation of oligomeric polyadenylated RNAs and their processing into mRNA. Mol. Cell. Biol. **4**:722–729.
2. **Andrews, E., and D. DiMaio.** Unpublished observations.
3. **Aviv, H., and P. Leder.** 1972. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. Proc. Natl. Acad. Sci. USA **69**:1408–1412.
4. **Baker, C.** 1989. Bovine papillomavirus type 1 transcription, p. 91–112. *In* H. Pfister (ed.), Papillomaviruses and human cancer. CRC Press, Boca Raton, Fla.
5. **Baker, C., and P. Howley.** 1987. Differential promoter utilization by the bovine papillomavirus in transformed cells and in productively infected wart tissues. EMBO J. **6**:1027–1035.
6. **Baker, C., and J. S. Noe.** 1989. Transcriptional termination between bovine papillomavirus type 1 (BPV-1) early and late polyadenylation sites blocks late transcription in BPV-1-transformed cells. J. Virol. **63**:3529–3534.
7. **Brown, P. H., L. S. Tiley, and B. R. Cullen.** 1991. Effect of RNA secondary structure on polyadenylation site selection. Genes Dev. **5**:1277–1284.
8. **Burnett, S., J. Moreno-Lopez, and U. Pettersson.** 1988. A novel spontaneous mutation of the bovine papillomavirus-1 genome. Plasmid **20**:61–74.
9. **Chirgwin, J. M., A. E. Przybyla, R. J. MacDonald, and W. J. Rutter.** 1979. Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry **18**:5294–5301.
10. **Connelly, S., and J. L. Manley.** 1988. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes Dev. **2**:440–452.
11. **Denome, R. M., and C. N. Cole.** 1988. Patterns of polyadenylation site selection in gene constructs containing multiple polyadenylation signals. Mol. Cell. Biol. **8**:4829–4839.
12. **DeZazzo, J. D., and M. J. Imperiale.** 1989. Sequences upstream of AAUAAA influence poly(A) site selection in a complex transcription unit. Mol. Cell. Biol. **9**:4951–4961.
13. **DiMaio, D.** 1986. Nonsense mutation in open reading frame E2 of bovine papillomavirus DNA. J. Virol. **57**:475–480.
14. **DiMaio, D., R. H. Treisman, and T. Maniatis.** 1982. Bovine papilloma virus vector that propagates as a plasmid in both mouse and bacterial cells. Proc. Natl. Acad. Sci. USA **79**:4030–4034.
15. **Doniger, J., C. D. Woodworth, and J. A. DiPaolo.** 1990. Alternative HPV 16 early message termination sites. UCLA Symp. Mol. Cell. Biol. **124**:519–522.
16. **Engel, L. W., C. A. Heilman, and P. M. Howley.** 1983. Transcriptional organization of bovine papillomavirus type 1. J. Virol. **47**:516–528.
17. **Fitzgerald, M., and T. Shenk.** 1981. The sequence 5′-AAUAAA-3′ forms part of the recognition site for polyadenylation of late SV40 mRNAs. Cell **24**:251–260.
18. **Frohman, M. A., M. K. Dush, and G. R. Martin.** 1988. Rapid amplification of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA **85**:8998–9002.
19. **Furth, P. A., and C. C. Baker.** 1991. An element in the bovine papillomavirus late 3′ untranslated region reduces polyadenylated cytoplasmic RNA levels. J. Virol. **65**:5806–5812.
20. **Galli, G., J. W. Guise, M. A. McDevitt, P. W. Tucker, and J. R. Nevins.** 1987. Relative position and strengths of poly(A) sites as well as transcription termination are critical to membranes versus secreted μ-chain expression during B-cell development. Genes Dev. **1**:471–481.
21. **Glisin, V., R. Crkvenjakov, and C. Byus.** 1973. Ribonucleic acid isolated by cesium chloride centrifugation. Biochemistry **13**:2633–2637.
22. **Grundstein, M., and D. Hogness.** 1975. Colony hybridization: a method for the isolation of cloned DNA's that contain a specific gene. Proc. Natl. Acad. Sci. USA **72**:3961–3965.
23. **Heilman, C. A., L. Engel, D. R. Lowy, and P. M. Howley.** 1982. Virus-specific transcription in bovine papillomavirus-transformed mouse cells. Virology **119**:22–34.

23a. **Kunkel, T.** 1985. Rapid and efficient site-specific mutagenesis without phenotypic selection. Proc. Natl. Acad. Sci. USA **82**:488–492.

24. **Lee, S., and E. Andrews.** Unpublished observations.
25. **Leff, S. E., M. G. Rosenfeld, and R. M. Evans.** 1986. Complex transcriptional units: diversity in gene expression by alternative RNA processing. Annu. Rev. Biochem. **55**:1091–1118.
26. **Logan, J., E. Falk-Pederson, J. E. Darnell, and T. Shenk.** 1987. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse β^{maj}-globin gene. Proc. Natl. Acad. Sci. USA **84**:8306–8310.
27. **Manley, J. L., H. Yu, and L. Ryner.** 1985. RNA sequence containing hexanucleotide AAUAAA directs efficient mRNA polyadenylation in vitro. Mol. Cell. Biol. **5**:373–379.
28. **Melton, D. A., P. A. Krieg, M. R. Rebagliati, T. Maniatis, K. Zinn, and M. R. Green.** 1984. Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. Nucleic Acids Res. **12**:7035–7056.
29. **Orkin, S., T. Cheng, S. Antonarakis, and H. Kazazian.** 1985. Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human β-globin gene. EMBO J. **4**:453–456.
30. **Proudfoot, N.** 1991. Poly(A) signals. Cell **64**:671–674.
31. **Russo, P., W.-Z. Li, D. M. Hampsey, K. S. Zaret, and F. Sherman.** 1991. Distinct cis-acting signals enhance 3′ endpoint formation of *CYC1* mRNA in the yeast *Saccharomyces cerevisiae*. EMBO J. **10**:563–571.
32. **Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. Mullis, and H. A. Erlich.** 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239**:487–491.
33. **Sarver, N., J. C. Byrne, and P. M. Howley.** 1982. Transformation and replication in mouse cells of a bovine papillomavirus-pML2 plasmid vector that can be rescued in bacteria. Proc. Natl. Acad. Sci. USA **79**:7147–7151.
34. **Scharf, S. J., G. T. Horn, and H. A. Erlich.** 1986. Direct cloning and sequence analysis of enzymatically amplified genomic sequences. Science **233**:1076–1078.
35. **Stoler, M. H., and T. R. Broker.** 1986. In situ hybridization detection of human papillomavirus DNAs and messenger RNAs in genital condylomas and a cervical carcinoma. Hum. Pathol. **17**:1250–1258.
36. **Whitelaw, E., and N. Proudfoot.** 1986. α-Thalassemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3′ end processing in the human α-2 globin gene. EMBO J. **5**:2915–2922.
37. **Wickens, M.** 1990. How the messenger got its tail: addition of poly(A) in the nucleus. Trends Biochem Sci. **15**:277–281.
38. **Wickens, M., and P. Stephenson.** 1984. Role of the conserved AAUAAA sequence: four AAUAAA point mutations prevent messenger RNA 3′ end formation. Science **226**:1045–1051.
39. **Yang, Y. C., H. Okayama, and P. M. Howley.** 1985. Bovine papillomavirus contains multiple transforming genes. Proc. Natl. Acad. Sci. USA **82**:1030–1034.

Pergamon

Chemical Engineering Science, Vol. 51, No. 23, pp. 5091–5102, 1996

0009-2509/96 $15.00 + 0.00

PII: S0009-2509(96)00288-6

DIRECTED EVOLUTION: CREATING BIOCATALYSTS FOR THE FUTURE

FRANCES H. ARNOLD
Division of Chemistry and Chemical Engineering 210-41, California Institute of Technology, Pasadena, CA91125, U.S.A.

(*Received* 10 *October* 1995; *accepted* 28 *January* 1996)

Abstract—An effective approach to engineering new enzymes is to direct their evolution *in vitro*. By mimicking key processes of Darwinian evolution in the test tube, the functions of enzymes can be explored free from the constraints of function within a living system. Efficient strategies for engineering new enzymes by multiple generations of random mutagenesis and recombination coupled with screening for improved variants have been developed. Our results with industrially important biocatalysts underscore the advantages of this 'evolutionary' approach to protein engineering. (From a talk presented at the first National Academy of Engineering "Frontiers of Engineering" Symposium, 21 September 1995.) Copyright © 1996 Elsevier Science Ltd

INTRODUCTION

Biological systems are the masters of chemical synthesis. The remarkable specificity of their catalysts, the enzymes, allows hundreds of reactions to proceed simultaneously inside the tiny reactor that is a living cell. Enzymes' ability to carry out complex chemical reactions, and to do so under very mild conditions with virtually no waste products, has earned them the admiration of chemists and biochemists. It is easy to envision that a future chemical industry sensitive to both energy needs and the environment could be modeled after these highly efficient chemical factories.

The molecules responsible for this remarkable performance are the enzymes. Enzymes are proteins, linear chains of typically hundreds of amino acids that fold up into unique, well-defined three-dimensional structures. The backbone of the polymer chain folds into a structure that is unique to the particular catalyst, as illustrated in Fig. 1(a) for the enzyme subtilisin. The enzyme's substrate (gray), the compound on which the reaction is catalysed, fits snugly into the substrate binding pocket. The enzyme positions specific catalytic amino acid side chains (red) where they can assist the chemical reaction to proceed. In Fig. 1(b) the structure of subtilisin showing its amino acid side chains illustrates the complexity of these molecular machines. This complexity allows enzymes to perform the truly impressive functions that support life and create new life. The result of considerable fine-tuning over eons of evolution, this complexity also makes it difficult to manipulate these structures to obtain new and interesting properties.

An enzyme is defined by a unique sequence of amino acids, which in turn is dictated by the organism's DNA code (the gene) and assembled in the cell (Fig. 2). This amino acid sequence determines how the chain folds and, ultimately, how the enzyme functions. By modifying the amino acid sequence, we can alter the enzyme's function—this field is known as *protein engineering*. Despite intense research into fundamental features governing protein folding and function, there are enormous gaps in our understanding of two critical processes: the relationship between sequence and structure and the relationship between structure and function. As a result, the rational design of new proteins by the classical 'reductionist' approach can be a frustrating exercise indeed. In this article I will introduce a new and highly effective approach to enzyme design and engineering that bypasses the need to understand these processes before embarking on a protein engineering project. But first I will explain why the enzymes provided by nature are not sufficient.

Chemical engineers who try to design real industrial processes using biological catalysts are constantly stymied by a simple fact: biological systems have evolved over billions of years to perform very specific biological functions and to do so within the context of a living organism. Some of the features required for function in a complex chemical network are undesirable when the catalyst is lifted out of context. Conversely, many of the properties we wish an enzyme would have clash with the needs of the organism, or at least were never required. The chemical engineer is hardly impressed by a catalyst whose inability to tolerate the most common of industrial conditions necessitates complicated hardware and reactors of the size of football fields. We need catalysts which are stable to high temperatures, can function in solvents other than water, tolerate wider ranges of pH,

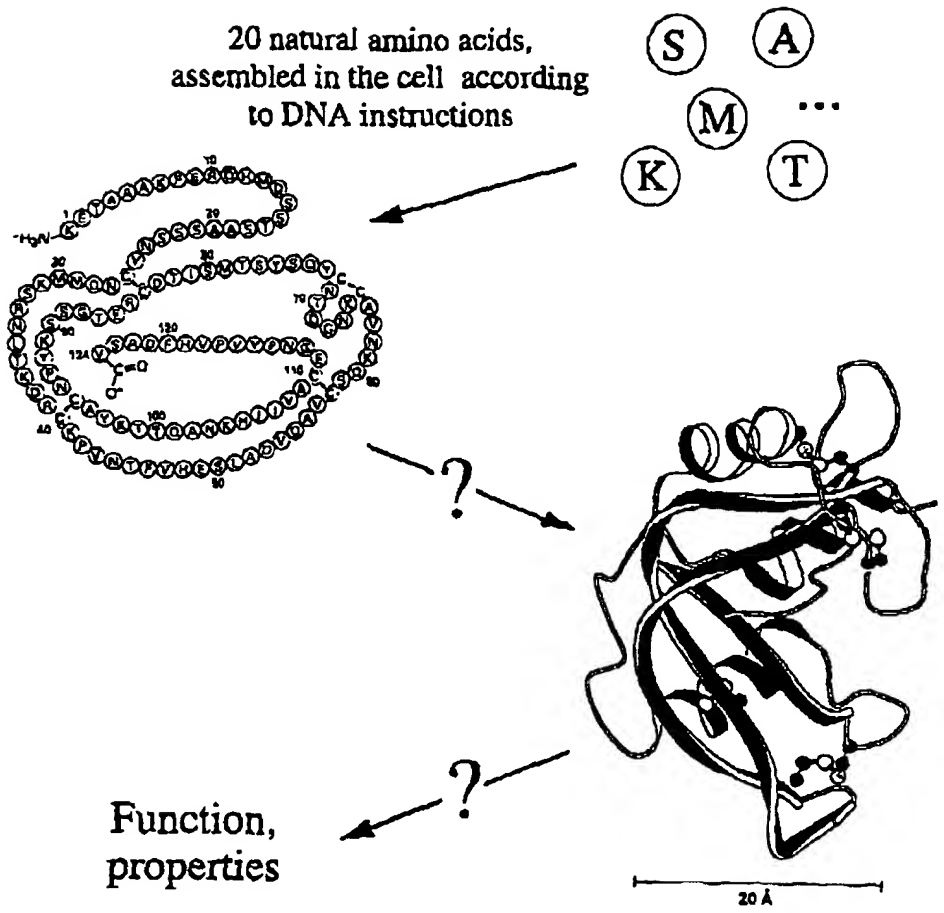

Fig. 2. Protein engineering involves the manipulation of protein structures and functions at the level of the amino acid (or DNA) sequence. Significant gaps in our understanding of the relationships between sequence, structure and function severely limit our ability to 'rationally design' new functions.

catalyse reactions on substrates not encountered in nature, and even catalyse new reactions not found in nature.

Many clues as to how to engineer better enzymes come from studying how nature has created enzymes. By studying the evolution of natural proteins, we have learned in fact that they are highly adaptable, constantly changing molecules, at least over evolutionary time scales. They can adapt to new environments and they can even take on new tasks. We know, for example, that many enzymes catalysing very different reactions have come about by divergent evolution from a common ancestral protein of the same general structure, acquiring diverse capabilities by processes of random mutation, recombination, and natural selection. For example, the versatile protein structure known as the α/β barrel diverged somewhere in the distant past to create a whole series of enzymes we know today (Reardon and Farber, 1995). The four enzymes shown in Fig. 3(a), for example, catalyse quite different reactions; their physical properties and amino acid sequences are also quite disparate. It is useful to note that, while the barrel-like protein fold is highly conserved, the amino acid sequences and functions of these enzymes are not.

A fascinating *recent* example of enzyme evolution is the appearance of phosphotriesterase, an α/β barrel enzyme that hydrolyses, at diffusion-limited rates, pesticides and chemical warfare agents that have existed only for about 50 years. It has been suggested that this enzyme, discovered in a soil bacterium, evolved during the last 50 years from a related sequence identified in the common *E. coli* bacterium and now known as the 'phosphotriesterase homology protein' (Scanlan and Reid, 1995). The biological function of this latter protein is unknown.

We also know that enzymes of a given function (for example, all catalysing a particular step in a metabolic pathway) can exhibit widely different properties (stability, solubility, tolerance to pH, etc.), depending on where they are found. For example, the three glyceraldehyde phosphate dehydrogenase (GAPDH) enzymes listed in Fig. 3(b) have very similar three-dimensional structures; their sequences are less similar. We know

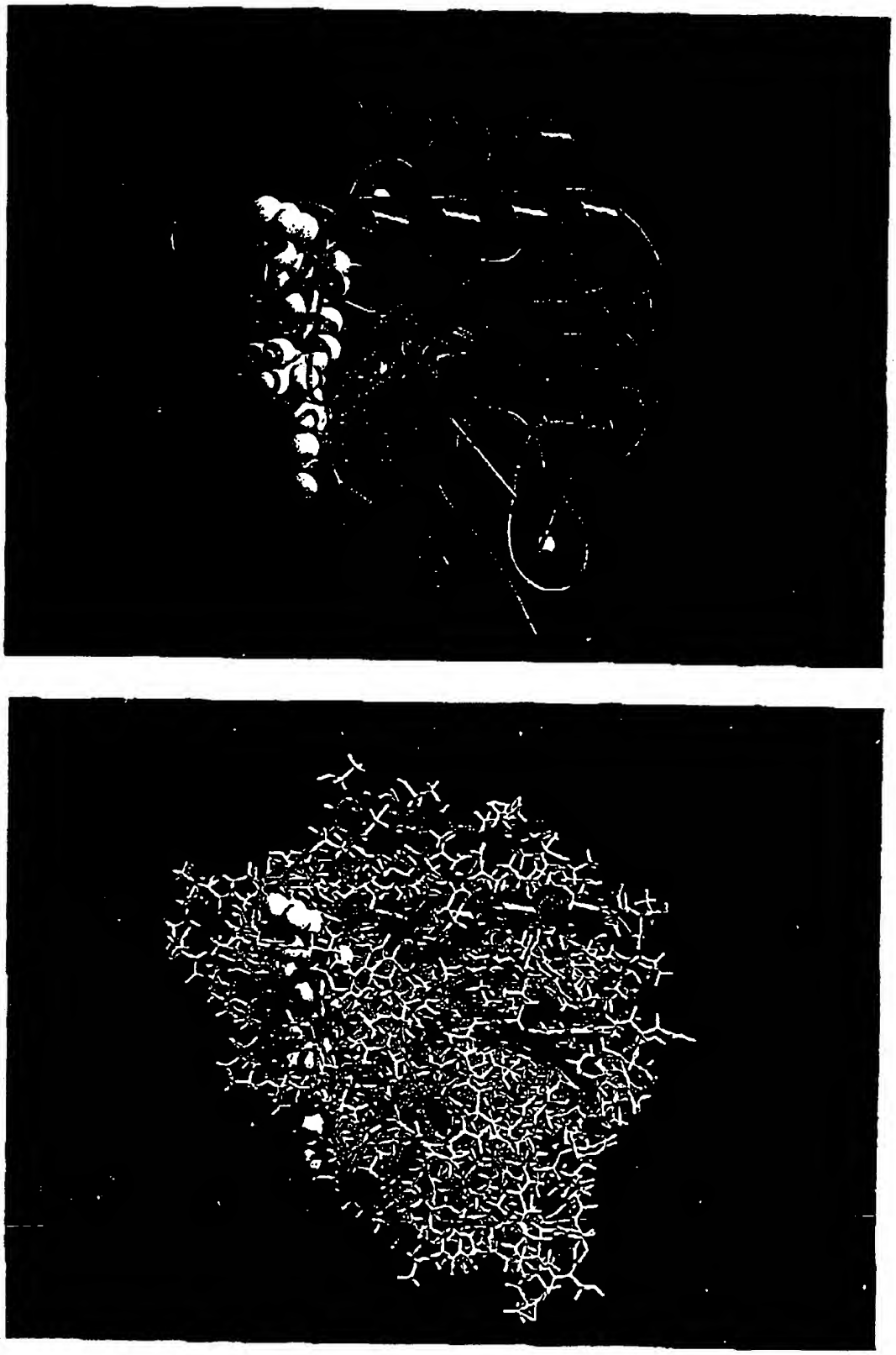

Fig. 1. The 275 amino acids of subtilisin E fold into a unique three-dimensional structure. (a) The backbone fold is represented here by a "ribbon" diagram, constructed from X-ray crystal structure coordinates (Dauter *et al.*, 1991) using the programs MolScript and Raster3D. Peptide substrate and two stabilizing calcium ions are shown in gray. Side chains of catalytic amino acid residues are shown in red. (b) Subtilisin structure showing the positions of the amino acid side chains (yellow).

Fig. 6 Molecular model of subtilisin E showing the 12 amino acid substitutions that increase enzyme activity in DMF (You and Arnold, 1996). Yellow amino acids were accumulated during screening for enhanced specific enzyme activity (Chen and Arnold, 1993). Red amino acids were found during screening for total (expressed) enzyme activity (You and Arnold, 1996). Calcium ions and peptide substrate are shown in gray.

Fig. 10. Molecular model of the pNB esterase showing positions of antibiotic *p*-nitrobenzyl ester substrate (yellow), catalytic residues (red), and six beneficial mutations accumulated during directed evolution (orange) (Moore and Arnold, 1996).

that th
a long
Therm
temper
enzym

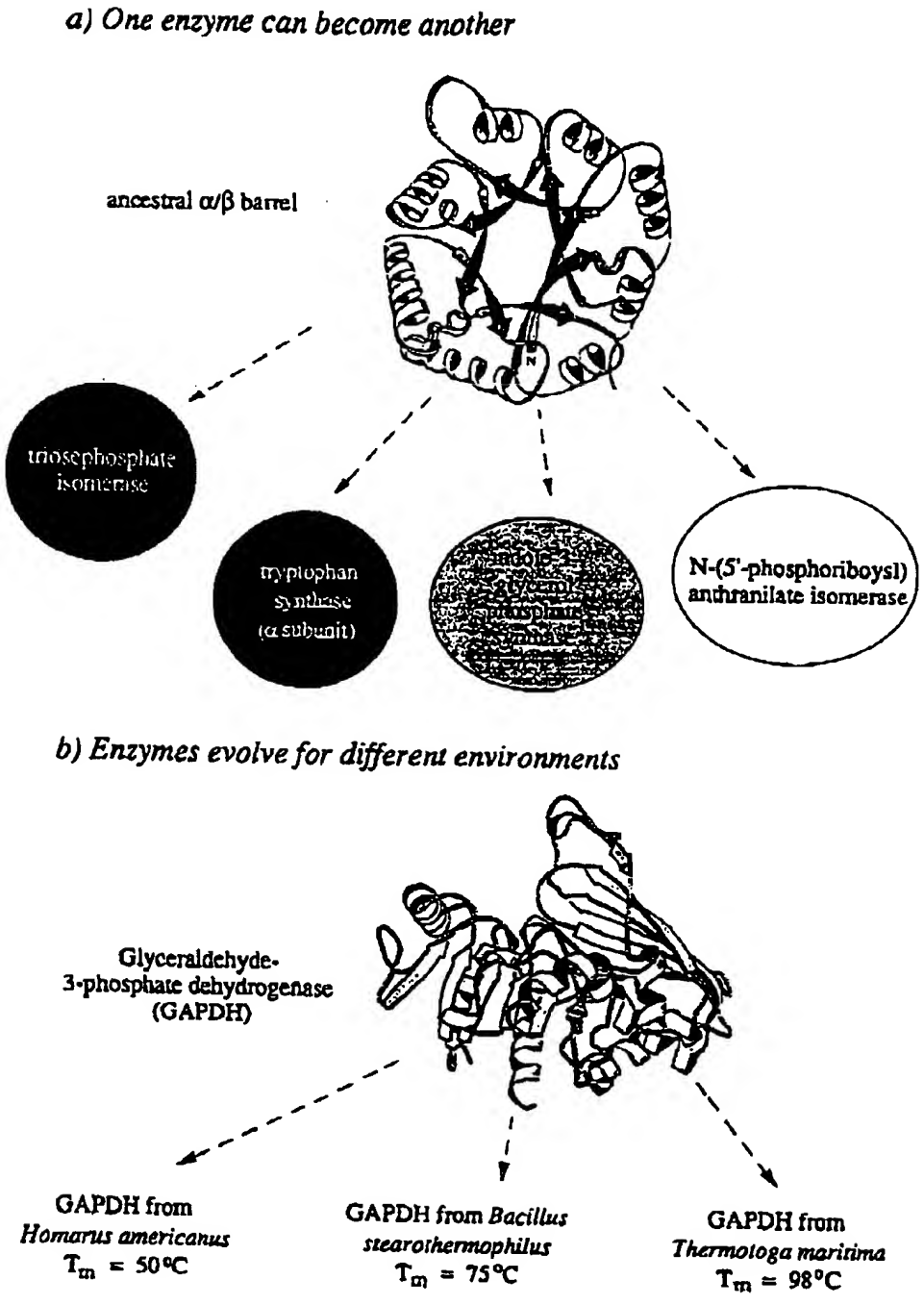

Fig. 3. Structure is conserved during evolution, while amino acid sequences and specific functions are often not. (a) The α/β barrel enzymes indicated appear to have evolved from a common ancestral α/β barrel protein. (b) Three GAPDH enzymes isolated from different organisms have very similar structures, but quite different stabilities and amino acid sequences (Buehner *et al.*, 1974; Skarzynski *et al.*, 1987; Korndoerfer *et al.*, 1995).

that they, too, diverged from some common ancestor a long time ago to occupy their current niches. The *Thermotoga maritima* bacterium thrives at very high temperatures in ocean thermal vents; consequently, its enzymes can tolerate much higher temperatures than the analogous enzymes from an organism which grows under less extreme conditions, such as *B. stearothermophilis*. The *Thermotoga* protein unfolds at 98°C, while the same enzyme from the American lobster unfolds at only 50°C. As with the α/β barrel

enzymes, the structural fold of the GAPDHs is highly conserved, while the detailed amino acid sequences and specific properties are not.

DIRECTED EVOLUTION: EXPLORING NEW FUTURES

The explosion of tools that has come out of molecular biology during the last 20 years has made it possible for us to consider 'evolving' the components of biological systems—DNA, RNA and proteins—for features never required in nature. We can both speed up the rate and channel the direction of evolution by controlling mutagenesis the rate and types of changes made—and the accompanying 'selection' pressures. As a result, processes that would take millions of years in nature can in principle be accomplished during the time scale of a Ph.D. thesis. By uncoupling the enzymes from the constraints of function within a living system, we can step into and explore a variety of futures, futures that can include novel environments (evolution in a sea of methanol instead of water?) or even entirely new functions (enzymes to break down hazardous chemicals?). We can explore questions such as 'can one catalytic activity become another, and how?' Furthermore, by evolving new functions and thereby new solutions to molecular design problems, we learn things about these amazing molecular machines that might never be revealed if we were to study only those that exist in nature.

The possibilities for biotechnology are especially exciting. Directed evolution is a very practical approach to tailor-making enzymes for a wide range of applications. In addition to building enzymes with new features and functions, we can explore important questions such as 'how might an enzyme change its sequence and properties to break down or evade a drug?' We could conceivably anticipate in laboratory experiments what might happen to drug resistances in nature. In directed evolution experiments we could also tune enzymes to function optimally under conditions specified by us, rather than the context of the living organism in which it evolved. New enzymes could be evolved to carry out reactions never required by living organisms.

DEVELOPING A WORKING STRATEGY FOR DIRECTED ENZYME EVOLUTION

In a directed evolution experiment, we first generate a library of many different possible 'solutions' to a problem. The next step is to find the *correct* solution(s), enzymes that exhibit the desired property. A conceptual challenge comes in planning how to create this library of solutions. The number of possible enzymes one can make is so vast that an exploration of their functions must be carefully guided in order to avoid becoming hopelessly lost. A typical enzyme is a linear polymer of 300 amino acids. With 20 possible amino acids at each position in the chain, there are 20^{300} possible different linear combinations. If even only a small fraction—say, 1 in 10^{10}—of all these sequences folds into a well-defined three-dimensional structure, there are still more structured proteins than there are atoms in the universe! (Note that even in three billion years, nature has not had a chance to explore but a tiny fraction of the possibilities. This also means that there are very exciting possibilities for future evolution, including evolution in the test tube.) Because a random sampling of amino acid sequences is unlikely to lead to the desired protein, we must begin our exploration by starting from a point that we hope is close to where we want to be—an enzyme that approximates what we want, but is not ideal. Then we evolve it, by accumulating small changes, similar to what happens in nature.

Nature is very good at searching mutant libraries for useful solutions. Unfavorable mutations are winnowed out at the same time as beneficial mutations are amplified, by linking the organism's growth rate and reproductive success to the performance of its components. In this process of *selection*, those organisms which grow faster quickly dominate, allowing an efficient search of very large populations (10^9 or more for bacteria).

Unfortunately, many of the features that are of interest to us cannot be linked to the survival or growth of the host organism—the prerequisite to selection. Enzymes, for example, can tolerate a variety of environments that cannot sustain life, so that the organism dies long before the enzyme has a chance to 'show its stuff'. For most problems of practical interest, in fact, mutant enzyme libraries must be *screened* rather than selected, one enzyme at a time. That is, the enzyme variants must be tested individually (screened) for the property of interest. This unfortunate reality effectively limits the search for improvements to mutant libraries containing perhaps 10^4–10^6 variants, several orders of magnitude smaller than what one can search when survival depends on success.

The strategy for molecular evolution is then illustrated by calculating how many different sequences one can create by starting from a given enzyme and making a few amino acid substitutions, as shown in Table 1. While there are only 5700 possible *single* mutants of a 300 amino acid enzyme, there are still more than 30 billion different sequences that differ from the original enzyme at only three positions. While a rapid screen might be able to cover a large fraction of all single mutants, and even some significant fraction of all double mutants, screening would be unable to give more than a very sparse sampling of the enzymes with multiple amino acid substitutions. Unless a vast majority of the mutations led to the desired property, dealing with a library of multiple mutations would be an experiment based on wishful thinking! (As might be expected for a finely tuned molecular machine, most mutations are deleterious or at least neutral; beneficial mutations are generally rare. The frequency with which one can expect to find beneficial mutations will depend on the extent to

which the particular feature of interest has already been optimized: the pathways up the mountain necessarily decrease in number as the pinnacle is approached.) Because luck is generally not an acceptable basis for the success of an experiment, the search is effectively limited to proteins with sequences and therefore properties very similar to their parents. In addition, we must be able to tune the rate or mode of mutation to produce libraries of primarily single amino acid substitutions.

The principles and power of directed evolution are best illustrated with examples. The first example will be the evolution of an enzyme to function in a polar organic solvent. It is well known that subtilisin, which normally cuts up peptides and proteins by cleaving the peptide bonds linking the amino acids together, will also catalyse peptide bond formation. Peptide bond formation is favored in organic media, as water participates in unwanted side reactions as well as hydrolysis of the product. Subtilisin actually remains folded and reasonably stable in high concentrations of polar organic solvents such as dimethylformamide (DMF). Unfortunately, the *catalytic activity* is very low. There is no fundamental reason, however, why subtilisin could not function in DMF—the enzyme's unhappiness reflects a balance among a very large number of noncovalent interactions in the system—protein, solvent, substrates and products—a balance that is adversely affected when the protein is dissolved in a nonaqueous medium. Because these complex interactions are poorly understood, we could not address this problem by a rational design approach. We therefore took the 'irrational' approach and asked whether we could 'evolve' a subtilisin that would function well in DMF (Chen and Arnold, 1991, 1993; You and Arnold, 1996).

The arguments set out above led us to the strategy for directing the evolution of an enzyme to perform a new function (or, in this case of subtilisin, an old function but under new conditions) illustrated in Fig. 4. In comparison to the enzyme performing a function for which it is selected, peptide hydrolysis in aqueous media, the new job is performed poorly indeed. Subtilisin has not been selected for hydrolysis in DMF, and there is, not surprisingly, a great deal of room for improvement. Because it is feasible to search only those subtilisin mutants with one or two amino acid substitutions, we will create and screen a library of such mutants for progeny slightly better than their parent. *The screening method for identifying useful mutations should ensure that the expected small enhancements brought about mainly by single mutations can be measured.* Although these progenies will generally resemble their parents, after many generations new features can develop, such that the descendents can be quite different from their ancestor. *Therefore, the generation of new, useful enzymes also relies on having an effective strategy for accumulating many such small improvements.* One such strategy involves carrying out sequential generations of random mutagenesis on the gene (DNA sequence coding for the enzyme) to create a mutant library, coupled with screening of the resulting proteins. In each generation a single variant is chosen as the parent for the next generation, and sequential cycles allow the evolution of the desired features.

We implemented this strategy to evolve subtilisin to function in DMF. A powerful molecular biology tool, the polymerase chain reaction (PCR), was used to make millions of copies of the gene that codes for the natural, or wild-type enzyme. By carrying out this (enzymatic) reaction under sub-optimal conditions, we could introduce base substitutions randomly throughout the DNA at a controllable rate. At the end of this reaction we have millions of gene copies, most slightly different from the wild-type one. These genes are placed back into a circular double-stranded piece of DNA (a plasmid) that contains all the instructions the bacterial cells need to translate the DNA into protein. When the bacteria are transformed with these plasmids, we have millions of individual chemical factories, each producing a different variant of the original enzyme.

Table 1. The molecular evolution 'number problem'

No. of amino acid changes	No. of possible variants
0	1
1	5700
2	16,190,850
3	30,557,530,900
4	43,109,036,717,100
5	48,489,044,499,400,000

Note: Starting with an enzyme of 300 amino acids, the number of sequences containing M amino acid substitutions is $(19)^M\ 300!/[(300 - M)!M!]$.

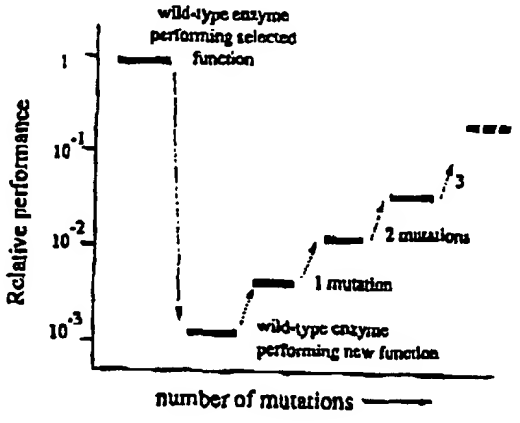

Fig. 4. A working strategy for directed enzyme evolution. The screening method should ensure that small enhancements brought about mainly by single mutations can be measured. The evolution of a new, useful enzyme requires an effective strategy for accumulating many such small improvements.

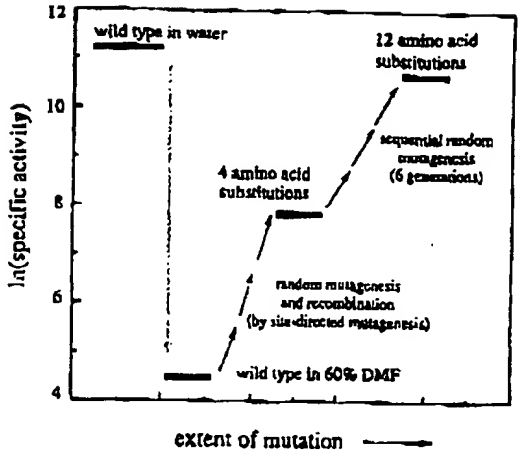

Fig 5. Results of directed evolution of subtilisin for activity in DMF by sequential generations of random mutagenesis and screening. The accumulation of 12 amino acid substitutions in sequential generations of random mutagenesis and screening resulted in an enzyme >500-fold more active than the wild-type enzyme in 60% DMF.

Next, the bacterial colony or colonies which produce a subtilisin that is more active in DMF must be found. In this early experiment our screening strategy was crude, but effective. Because subtilisin is secreted from the bacilli, the variants could be screened visually on nutrient plates containing a protein (casein), in the presence and absence of DMF. The active enzyme creates a visible 'halo' surrounding the bacterial colony whose size is proportional to the catalytic activity.* Variants with higher hydrolysis activity than wild-type on the DMF-containing plates could be identified from their bigger halos (Chen and Arnold, 1991).

The results of the directed evolution effort are summarized in Fig. 5. At first we identified three amino acid substitutions that individually improved the wild-type enzyme's activity several-fold. Using site-directed mutagenesis we combined those three with a fourth mutation reported to improve activity and stability in other subtilisins, to obtain a four amino acid variant about 40-fold more active than wild-type in 60% DMF (Chen and Arnold, 1991). Since the process of sequencing the genes of all the positive variants and then combining the mutations by site-directed mutagenesis was laborious, we decided to carry out sequential generations of random mutagenesis and screening, no longer stopping on the way to sequence the intermediates. Applying an additional six generations of mutagenesis and screening a few hundred colonies in each generation, we created an enzyme that is more than 500-fold more active in 60% DMF than the wild-type subtilisin E (You and Arnold, 1996). This enzyme exhibits substantial activity even in 85% DMF. The whole process was surprisingly rapid: a total of only about 10,000 colonies were screened to obtain a huge improvement in catalytic activity.

The gene for the final evolved enzyme was sequenced to determine the amino acid substitutions that allowed this enzyme to recover its activity in DMF. Of 275 amino acids, 12 were altered; their positions are indicated in Fig. 6. Although the DNA substitutions are targeted randomly throughout the entire subtilisin gene sequence, the amino acid substitutions that enhance catalytic activity are all positioned on the surface of the enzyme, surrounding the active site and substrate binding pocket. The majority are in evolutionarily variable loops that connect elements of conserved secondary structures (helices and sheets) (Chen and Arnold, 1993; You and Arnold, 1996). This information could of course be utilized in developing more 'rational' design strategies, including narrowing the sequences exposed to random mutagenesis in directed evolution.

Finally, it is worth noting that the resulting enzyme is indeed a far more efficient catalyst than wild-type subtilisin for the polymerization of amino acids. This evolved enzyme can catalyse, for example, the formation of poly-L-methionine starting from a racemic mixture of methionine methyl ester. The evolved enzyme allows the synthesis of significantly longer polymers and at much higher yields than the native enzyme in 60–70% DMF (Zhao, H. unpublished results).

The advantage of directed evolution over site-directed mutagenesis is clear: the same amount of effort could support the construction and screening of at most a few dozen variants with mutations directed to specific locations. Without a clear mechanism, it would be difficult indeed to pinpoint 12 amino acid substitutions that enhance catalytic activity in DMF. Even then, single site-directed mutations would have to be accumulated to create a useful enzyme, itself a substantial mutagenesis effort involving trial and error to find optimal combinations.

The most attractive feature of the evolutionary strategy outlined in Fig. 4 is its simplicity. It is possible, however, that this simple 'up-hill climb' approach is not an optimal approach to the evolution of a particular enzyme. There are obviously a great number of pathways possible for the evolution of a protein, and each choice of parent for the next generation represents an irreversible step along one particular pathway. What would happen if we simply repeated the experiment? Depending on which pathway was chosen or which mutation happened to be found first, the enzyme could end up on a local optimum, unable to evolve further. This approach may also appear slow: improvements are small in each step and necessarily become harder to find the closer the enzyme gets to an optimum.

*Because halo size also depends on enzyme expression level, enzyme diffusion and colony size, it is useful for a 'rough cut'. Positives were confirmed by a second level of screening in liquid culture (Chen and Arnold, 1993).

SEX IN THE TEST TUBE

An alternative directed evolution strategy we have recently explored incorporates some important advantages attributed to sex in the evolutionary process. Gene recombination, the cutting and pasting of whole genes or pieces of genes, can significantly increase the speed of molecular evolution by rapidly accumulating beneficial mutations and providing a mechanism to remove deleterious ones. To incorporate recombination into directed evolution, we randomly recombine genes with positive mutations. A search for better *combinations* of mutations completes a generation of directed evolution.

We have tested this new 'sexual' approach by creating an enzyme that efficiently catalyses the hydrolysis of the *p*-nitrobenzyl (pNB) ester of a *β*-lactam antibiotic in the presence of DMF. The pNB protecting group is often used during the large-scale synthesis of cephalosporin-type antibiotics. Its selective removal presents problems, however, particularly for recovery and disposal of the zinc catalyst and the large amounts of organic solvents used. Therefore, a major pharmaceutical company devoted significant effort some years ago to finding an enzyme that would perform this selective hydrolysis reaction (Brannon *et al.*, 1976; Zock *et al.*, 1994). An enzyme with some activity towards pNB ester hydrolysis was identified by screening a large number of microorganisms, but the enzyme's low activity, especially in the solvents required to solubilize these materials, made it a *poor* competitor to the classical chemical catalyst.

We were challenged in 1994 to evolve a pNB esterase with much higher activity, particularly in the presence of the polar organic solvents required to achieve high substrate solubility. We had two reasons to believe that this could be done. First, the enzyme's natural function and, therefore, natural substrates are unknown, but they are unlikely to be the antibiotic pNB esters. Second, the natural enzyme's activity is very sensitive to organic solvents. Because these features were never required in the enzyme's natural setting, we could expect considerable improvement through directed evolution.

The wild-type esterase is not secreted by the *E. coli* cells in which it is made, nor does it carry out a reaction that is easily measured. Thus, we had to develop screening strategies more sophisticated than those used for the subtilisin. The *p*-nitrobenzyl ester hydrolysis reaction is assayed laboriously by high performance liquid chromatography, a method unsuitable for screening tens of thousands of colonies. We therefore devised a rapid screening assay using a similar, but not identical, *p*-nitrophenyl ester substrate, in order to have an easy-to-read colorimetric signal. The screening reactions could then be carried out in the 96 wells of a plastic microtiter plate, using an automatic spectrophotometer to read and analyse the absorbance in all 96 wells at once.

Using this rapid assay to screen about a thousand colonies per generation, we completed several sequential cycles of random PCR mutagenesis and screening, as illustrated in Fig. 7 (Moore and Arnold, 1996). After four generations, the enzyme's specific activity in 15% DMF had improved 15-fold. In the fourth generation, we collected not one, but 64 different clones, some of which were better than the parent, and many of which were not. The purpose for this was two-fold. First we wanted to make sure that our screening strategy was working properly to give us an enzyme that would catalyse the desired *p*-nitrobenzyl hydrolysis reaction, not only the colorimetric *p*-nitrophenyl screening reaction. The activities of each of the 64 clones in both reactions are compared in Fig. 8.

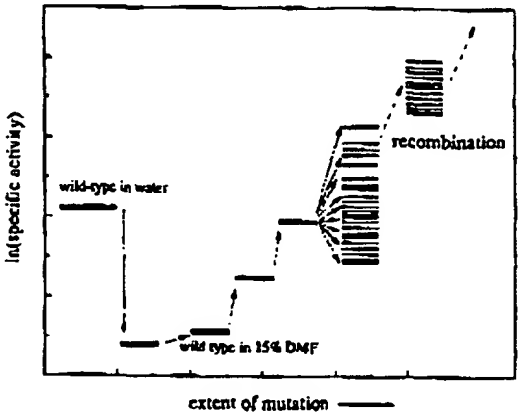

Fig. 7. Directed evolution of pNB esterase in 15% DMF involved four generations of random mutagenesis and screening, followed by one round of recombination of the five best genes from generation 4. The best variant obtained after four generations is 15-fold more active than wild-type. The best variant from screening 400 colonies of the recombination pool is ~30-fold more active than wild-type.

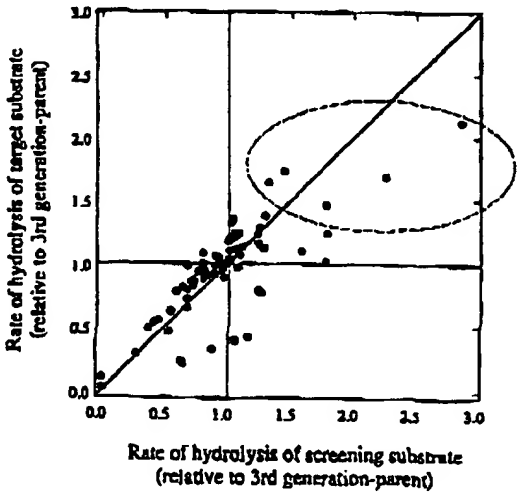

Fig. 8. Comparison of activities on target (p-nitrobenzyl) and screening (p-nitrophenyl) substrate of 64 pNB esterase variants isolated after fourth generation of random mutagenesis and screening, relative to parent enzyme from the third generation. The five most active variants (inside oval) were pooled for random recombination (see Fig. 9).

If the screening reaction perfectly mimicked the desired reaction, all the points would lie on the 45° line. Although somewhat scattered, there is none the less a reasonable correlation: the rapid screen provides an indication of evolution of the desired activity that is acceptable for making a rough cut of positive clones.

The second reason for studying this group of variants was to test the alternate, sexual approach for accumulating effective mutations. We thus collected the five best mutants, those in the dotted oval in Fig. 8, and recombined them using a 'sexual' PCR method recently described by Stemmer (1994a, b). How the genes are randomly recombined is shown schematically in Fig. 9(a). The genes are pooled in the test tube and fragmented with an enzyme that cuts the DNA at random positions. In Fig. 9(b), the polyacrylamide electrophoresis gel that separates the DNA fragments by length shows that the DNA has been digested into a smear of different-sized pieces. We collected the fragments 200–300 base pairs in length by extracting the DNA from the appropriate piece of gel. The full-length gene can be reassembled from this pool of random fragments, again using the PCR technology, to create a new gene library in which the mutations were present in their different possible combinations. These reassembled, recombined genes were inserted back into the plasmid and expressed in the *E. coli*. The best of those recombined genes were identified, as before, by screening the enzymes they code for and produce in the microorganisms.

Screening only ~400 colonies yielded eight clones with activity significantly greater than the best of the five parents—this yield of positives is at least 20-fold higher than we found by screening the genes with point mutations alone (typically 1/1500). Recombination can enhance directed evolution by making use of the information present in a population of improved enzymes produced by mutagenesis and screening, information that would otherwise be discarded. Thus far, we have improved the enzyme's specific activity towards the antibiotic substrate more than 30-fold in 15% DMF. The total expressed activity is at least 50-fold greater than the original system we started with.

Sequencing of the genes coding for improved enzymes once again allowed us to identify the amino acid substitutions responsible for the observed improvements in catalytic performance. Six effective mutations are illustrated in Fig. 10, on a model of the pNB esterase developed from the X-ray crystal structure of a homologous enzyme (Moore and Arnold, 1996). As for the case of subtilisin, most of the mutations are at or near the solvent-accessible surface. Only one of the six is deeply buried. In contrast to subtilisin, however, none of the effective amino acid substitutions lie in segments of the esterase predicted to interact directly with the bound substrate. It is possible that the homology modeling yielded an incorrect structure, and the mutations do interact with the *p*-nitrobenzyl substrate. Or, it may be that the amino acid substitutions sampled at positions adjacent to the substrate were all deleterious, and small improvements were only obtained by altering amino acids further away. In any case, the mechanism(s) by which these amino acid substitutions enhance the catalytic activity of the evolved pNB esterases are subtle and would have been very difficult to predict in advance.

CONCLUSIONS

The directed evolution approach clearly allows us to engineer enzymes with novel functions and features. In contrast to 'rational' design approaches, directed evolution can be applied even when very little is known about an enzyme's structure or catalytic mechanism. Since the vast majority of proteins remain largely uncharacterized, this marks a huge advantage for the evolutionary methods. This approach, because it allows us to explore novel solutions to protein design problems, also promises to teach us a great deal about protein structure and function.

Future research in directed evolution will include development of large-scale screening methods, so that efficient searches of large mutant libraries can be performed. The construction of optimized mutant libraries will also decrease the need for screening. In addition to streamlining efforts to 'tune' enzymes, these improvements will allow larger leaps—such as the evolution of new catalytic activities—to take place. Significant improvements in the ease and power of directed evolution will also come from optimizing the search strategies. The many similarities to optimization problems in other fields make this a fertile ground for collaborative efforts among theoreticians and experimentalists from a wide range of engineering disciplines.

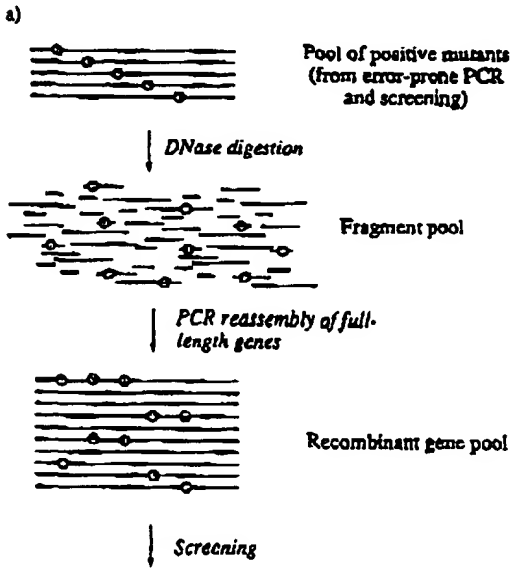

Fig. 9. Recombination of mutations by gene shuffling. (a) 'Sexual' PCR method (Stemmer, 1994a, b) involves random digestion of the gene pool using DNase enzyme, followed by gene reassembly using PCR. Reassembled genes contain the different combinations of mutations.

Fig. 9. (
200–300

Acknowl
postdoct
especiall
This wo
search a
Biologic
Office o
Renewal

Buehner
Rossn
n-glyc
Biol. S
Brannor
esterif
micro
Chen, J
of ar

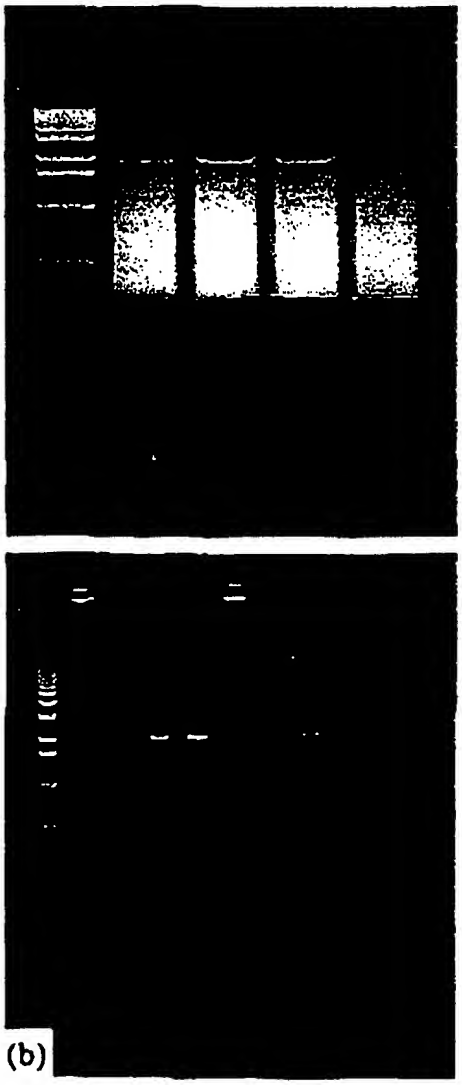

Fig. 9. (b) Top polyacrylamide electrophoresis gel shows the separation of the digested gene fragments by size. Fragments 200–300 base pairs long were recovered by extracting the excised gel segment. These were reassembled into the full-length gene (bottom gel, lanes 3–5 and 6–8). First lane on left is a 'ladder' of DNA of known molecular weights.

Acknowledgements—I am grateful to all the students and postdoctoral researchers who have contributed to this effort, especially Jeffrey Moore, Kevin Chen and Huimin Zhao. This work was supported by the U.S. Office of Naval Research and the U.S. Department of Energy's program in Biological and Chemical Technologies Research within the Office of Industrial Technologies, Energy Efficiency and Renewables.

REFRENCES

Buehner, M., Ford, G. C., Moras, D., Olsen, K. W. and Rossmann, M. G., 1974. Three-dimensional structure of D-glyceraldehyde-3-phosphate dehydrogenase. *J. Mol. Biol.* **90**, 25–49.

Brannon, D. R., Mabe, J. A. and Fukuda, D. S., 1976. De-esterification of cephalosporin para-nitrobenzyl esters by microbial enzymes. *J. Antibiotics* **29**, 121–124.

Chen, K. and Arnold, F. H., 1993. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 5618–5622.

Chen, K. and Arnold, F. H., 1991. Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin E in polar organic media. *Biotechnology* **9**, 1073–1077.

Dauter, Z., Betzel, C., Genov, N., Pipon, N. and Wilson, K. S., 1991. Complex between the subtilisin from a mesophilic bacterium and the leech inhibitor eglin-C. *Acta Crystallogr.* **B47**, 707–730.

Korndoerfer, I., Steipe, B., Huber, R., Tomschy, A. and Jaenicke, R., 1995. The crystal structure of hologlyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima* at 2.5 Å resolution. *J. Mol. Biol.* **246**, 511–521.

Moore, J. and Arnold, F. H., 1996. Directed evolution of a *p*-nitrobenzyl esterase for aqueous-organic solvents. *Nature Biotechnol.* **14**, 458–467.

Reardon, D. and Farber, G. K., 1995. Protein motifs 4. The structure and evolution of α/β barrel proteins. *FASEB J.* **9**, 497–503.

Scanlan, T. S. and Reid, R. C., 1995, Evolution in action. *Chem. Biol.* 2, 71–75.

Skarzynski, T., Moody, P. C. E. and Wonacott, A. J., 1987, Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus* at 1.8 Å resolution. *J. Mol. Biol.* 193, 171–187.

Stemmer, W. P. C., 1994a, DNA shuffling by random mutagenesis and reassembly: *in vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* 91, 10,747–10,751.

Stemmer, W. P. C., 1994b, Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* 340, 389–391.

You, L. and Arnold, F. H., 1996, Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide. *Protein Engng* 9, 77–83.

Zock, J., Cantwell, C., Swartling, J., Hodges, R., Pohl, T., Sutton, K., Rosteck, P. Jr., McGilvray, D. and Queener, S., 1994, The *Bacillus subtilis* pnbA gene encoding *p*-nitrobenzyl esterase: cloning, sequence and high-level expression in *E. coli*. *Gene* 151, 37–43.

The rec
(e.g. vac
a cell-di
the host
ccss-sca
(Middel
mechan
used me
and Ku
Num
as majc
homoge
Marcus
pingem
1981), t
(pressur
shear (A
elling a
they ha
homoge
simplifi
informa
a homo
and flo
previou
The
pressur
Gaulin
ted wit
remove
is used

*Cor